doi: 00.0000/jmbi.0000.0000 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (0000) 00, 1–20

JMB

Review

Studying Genomes Through the Aeons: Protein Families, Pseudogenes and Proteome Evolution

Paul M. Harrison and Mark Gerstein*

Department of Molecular Biophysics and Biochemistry Yale University, 266 Whitney Avenue, P.O. Box 208114, New Haven, CT 06520-8114, USA

Protein families can be used to understand many aspects of genomes, both their "live" and their "dead" parts (i.e. genes and pseudogenes). Surveys of genomes have revealed that, in every organism, there are always a few large families and many small ones, with the overall distribution following a power-law. This commonality is equally true for both genes and pseudogenes, and exists despite the fact that the specific families that are enlarged differ greatly between organisms. Furthermore, because of family structure there is great redundancy in proteomes, a fact linked to the small size of the minimal, indispensable sub-proteome for each organism and the large number of dispensable genes. Pseudogenes in prokaryotes represent families that are in the process of being dispensed with. In particular, the genome sequences of certain pathogenic bacteria (*Mycobacterium leprae*, *Yersinia pestis* and *Rickettsia prowazekii*) show how an organism can undergo reductive evolution on a large-scale (i.e. the dying out of families) as a result of niche change. There appears to be less pressure to delete pseudogenes in eukaryotes. These can be divided into two varieties, duplicated and processed, where the latter involves reverse transcription from an mRNA intermediate. We discuss these collectively in yeast, worm, fly, and human. The fly has few pseudogenes because of its high rate of genomic DNA deletion. In the other three organisms, the distribution of pseudogenes on the chromosome and amongst different families is highly non-uniform. Pseudogenes tend not to occur in the middle of chromosome arms and to be associated with lineage-specific (as opposed to highly conserved) families that have environmental-response functions. This may be because, rather than being dead, they may often form a reservoir of diverse "extra parts" that can be resurrected to help an organism adapt to its surroundings. In yeast, there may be a novel mechanism involving the [PSI +] prion that potentially enables this resurrection. In worm, the pseudogenes tend to arise out of families (particularly chemoreceptors) that are greatly expanded in it compared to the fly. The human genome stands out in having many processed pseudogenes. These have a character very different from those of the duplicated variety, essentially just representing random insertions. Thus, their occurrence tends to be roughly in proportion to the amount of mRNA for a particular protein and to reflect the extent of the intergenic sequences. Further information about pseudogenes is available at http://genecensus.org/pseudogene

*Corresponding author

Keywords: aeons; psuedogenes; proteome evolution

The complete or near-complete sequencing of the genomes of six eukaryotes (at the time of writing) and dozens of prokaryotes is enabling us to examine molecular evolution and diversity of proteins from a "whole-proteome" perspective. In the present review, we discuss various themes and

Abbreviations used: LINE, long interspersed nuclear element.

E-mail address of the corresponding author: mark.gerstein@yale.edu

0022-2836/02/not known0001–20 $35.00/0

issues in relation to proteome evolution, examining both the "live" and "dead" proteomes of specific genomes (all the proteins encoded by an organism and all the pseudogenes). We set the stage for discussion of pseudogene populations by surveying different issues relating to protein family redundancy in the live proteome, and how it evolves. In particular, we examine how such redundancy can be viewed in terms of partition into essential and dispensable sub-proteomes. Chiefly, then, we discuss the distribution of proteins and protein families in pseudogene populations for prokaryotes, and specifically for the eukaryotes yeast, worm, fly and human, and the implication of these dead or "dispensed-with" sequences for proteome evolution.

What is a protein family?

A protein family is usually defined as a group of sequences with an obvious evolutionary relationship, judged chiefly by protein sequence comparison, i.e. whose evolution can be studied readily at the sequence level. The definition of the threshold of similarity is arbitrary in practice and different degrees of protein sequence similarity are used depending on the context.[1–3] Membership of the same protein family is now commonly determined by the occurrence of a sequence motif indicative of sequence, structural and functional similarity, with integrated databases of such motifs used routinely in genome annotation.[4,5] There are now many databases that cluster protein sequences manually or automatically to varying degrees, at various levels of sequence and structural similarity (e.g. ProtoMap,[6] SYSTERS,[7] SCOP[8] and CATH[9]). As a higher level, a superfamily can then be described in terms of groups of families that have more distant similarity; they may have common evolutionary origin as judged by functional and structural similarities. (This is the definition used in the SCOP database.[8]) Different superfamilies can be grouped together if they have the same protein fold. Sometimes it is more appropriate to group families together into similar functional classes, e.g. the Gene Ontology database,[10] MIPS functional classification[11] or GenProtEc for *Escherichia coli*.[12] Although, usually, as for most of the work discussed below, robustness of results is reported for a range of sequence similarity cut-offs, there are a number of caveats in considering assignment of protein families and superfamilies to genomic data.[13,14] Firstly, such assignment procedures are biased towards larger families and superfamilies, in that sequence-searching procedures, such as the commonly used iterative program PSI-BLAST,[15] operate better for larger known families and are calibrated to search for larger families; secondly, for obvious reasons, gene prediction is more successful for them too.

Surveys of the "live" proteome

There has been extensive recent work on the counting of different levels of proteome parts: protein families, superfamilies and folds.[5,13,16–24] Initially, this work focused on prokaryotes, but is now shifting emphasis to the recently genomically sequenced eukaryotes. Surveys of protein fold and superfamily occurrence in microbial proteomes shows that a few folds predominate, whereas many folds occur only once; protein fold occurrences tend to rely on the prevalence of a single superfamily, although the rankings for these corresponding folds and superfamilies vary widely.[19] There are similar findings for the eukaryotes (Table 1).

Power-law distribution of protein family size in proteomes

Despite expansion and contraction in the size of individual protein families in proteomes, the redundancy in protein families appears to have a characteristic distribution common to viral, bacterial, archaeal and eukaryotic genomes.[3,16] An initial analysis of the distribution of the number of sequences in protein families *versus* their occurrence showed that the distribution for protein families in proteomes follows power-law behaviour (i.e. a linear relationship on a log–log plot), with a shallower slope for the relationship in the larger genomes[25] Huynen & Nimwegen[3] did a similar analysis for a larger number of microbial genomes and found that the power-law behaviour was maintained over a large range of sequence similarity thresholds used for clustering into families. They argued, using a simple probabilistic formalism, that the power-law distribution implies that gene duplications and deletions within gene families are largely dependent on one another. Other studies have shown that the distribution of the number of protein families and of protein folds in a proteome can be explained by simple evolutionary models that involve only duplication or the creation of new families or folds[22,26] An example of this power-law behaviour is illustrated in Figure 1 for families in the yeast proteome, and for protein folds and superfamilies.

Protein family redundancy in proteomes and its evolution in eukaryotes

The total number of protein domain sequence families, or functional diversity, appears to vary much less between organisms than overall proteome size. This is most striking in the eukaryotes.[2,27,28] For example, despite the wide variation in the number of annotated genes, the yeast, worm, fly and human proteomes seem to contain similarly sized subsets of the InterPro sequence domain database (851 for yeast; 1014 for worm; 1035 for fly; 1262 for human, at the time of writing).[5,27] The eukaryotic proteomes comprise

Table 1. Top-ranking protein superfamilies and folds in five eukaryote proteomes

Top-ranking superfamilies					Top-ranking folds				
Yeast	Worm	Fly	Mustard weed	Human	Yeast	Worm	Fly	Mustard weed	Human
P-loop NTP hydrolase (438)	P-loop NTP hydrolase (651)	C2H2 Zn finger (823)	P-loop NTP hydrolase (1282)	**C2H2 Zn finger, 7.37.1 (3424)**	P-loop NTP hydrolase, 3.32 (438)	Ig-like, 2.1 (1044)	*Ig-like, 2.1 (999)*	*α/α Superhelix, 1.111 (1475)*	**C2H2 Zn finger, 7.37 (3424)**
Protein kinase (133)	Ig (571)	P-loop NTP hydrolase (661)	Protein kinase (1070)	**Ig, 2.1.1 (1453)**	α/α Superhelix, 1.111 (195)	P-loop NTP hydrolase, 3.32 (651)	C2H2 Zn finger, 7.37 (823)	P-loop NTP hydrolase, 3.32 (1282)	**Ig-like, 2.1 (3034)**
WD-repeat (107)	Protein kinase (500)	*Ig, 2.1.1 (548)*	*Tetraticopeptide repeat, 1.111.8 (787)*	**P-loop NTP hydrolase, 3.32.1 (1229)**	Ferredoxin-like, 4.51 (154)	Protein kinase, 4.130 (500)	P-loop NTP hydrolase, 3.32 (661)	Protein kinase, 4.130 (1070)	**P-loop NTP hydrolase, 3.32 (1229)**
RNA-binding domain (104)	EGF/laminin (400)	EGF/laminin (330)	RNI-like (709)	**EGF/laminin, 7.3.9 (1083)**	Protein kinase, 4.130 (133)	Knottin, 7.3 (429)	α/α Superhelix, 1.111 (438)	Leucine-rich repeat, 3.9 (812)	**Knottin, 7.3 (1114)**
NADP-binding Rossmann fold (99)	C-type lectin (369)	Protein kinase (288)	RING finger (468)	Fibronectin type-III (817)	Seven-bladed β propeller, 2.64 (118)	α/α Superhelix, 1.111 (405)	Ferredoxin-like, 4.51 (357)	Ferredoxin-like, 6.51 (451)	α/α Superhelix, 1.111 (898)
ARM repeat (84)	Glucocorticoid receptor-like (349)	Spectrin repeat (268)	Homeodomain (461)	**Protein kinase, 4.130.1 (710)**	TIM barrel, 3.1 (114)	C-type lectin, 4.154 (369)	Knottin, 7.3 (345)	DNA/RNA-binding 3-Helical bundle, 1.4 (539)	**Protein kinase, 4.130 (710)**
DNA/RNA polymerises (59)	Nuclear receptor ligand-binding domain (284)	RNA-binding domain (257)	RNA-binding domain (426)	Cadherin (676)	RNase H, 3.50 (110)	Glucocorticoid receptor-like, 7.39 (349)	Protein kinase, 4.130 (288)	RING finger, 7.44 (468)	**Ferredoxin, 4.51 (655)**
Actin-like ATPase (56)	Homeodomain (263)	Trypsin-like serine protease (240)	NADP-binding Rossmann-fold domain (366)	**RNA-binding domain, 4.51.7 (517)**	NADP-binding Rossmann fold, 3.2 (99)	DNA/RNA-binding 3-helical bundle, 1.4 (329)	Spectrin repeat, 1.7 (272)	Seven-bladed β propeller, 2.64 (451)	**DNA/RNA-binding 3-helical bundle, 1.4 (510)**
Membrane all-α (54)	C2H2 Zn finger (255)	*Fibronectin type III, 2.1.2 (219)*	α/β Hydrolase (341)	**PH domain, 2.52.1 (415)**	DNA/RNA-binding 3-helical bundle, 1.4 (59)	Ferredoxin-like, 4.51 (301)	Trypsin-like serine protease, 2.44 (240)	TIM barrel, 3.1 (383)	**PH domain, 2.52 (415)**
Zn2/Cys6 DNA-binding domain (53)	α/β-Hydrolase (219)	*Cadherin, 2.1.6 (213)*	*ARM repeat, 1.111.1 (284)*	**Homeodomain, 1.4.1 (339)**	DNA/RNA polymerases, 5.8 (59)	Nuclear receptor ligand-binding domain, 1.116 (284)	Seven-bladed β-propeller, 2.64 (118)	NADP-binding Rossmann-fold domain, 3.2 (366)	Seven-bladed β propeller, 2.64 (394)

The Table shows the top-ranking folds in eukaryotes from SCOP. There is a pattern similar to that observed in prokaryotes.[19] In particular, for human, the prevalence of a fold tends to be due to a particular superfamily prevalence (superfamilies and folds in bold in the Table. Examples of folds that have multiple prevalent superfamilies are observed; examples for fly and mustard weed (*A. thaliana*) are in italics. Ig, immunoglobulin.

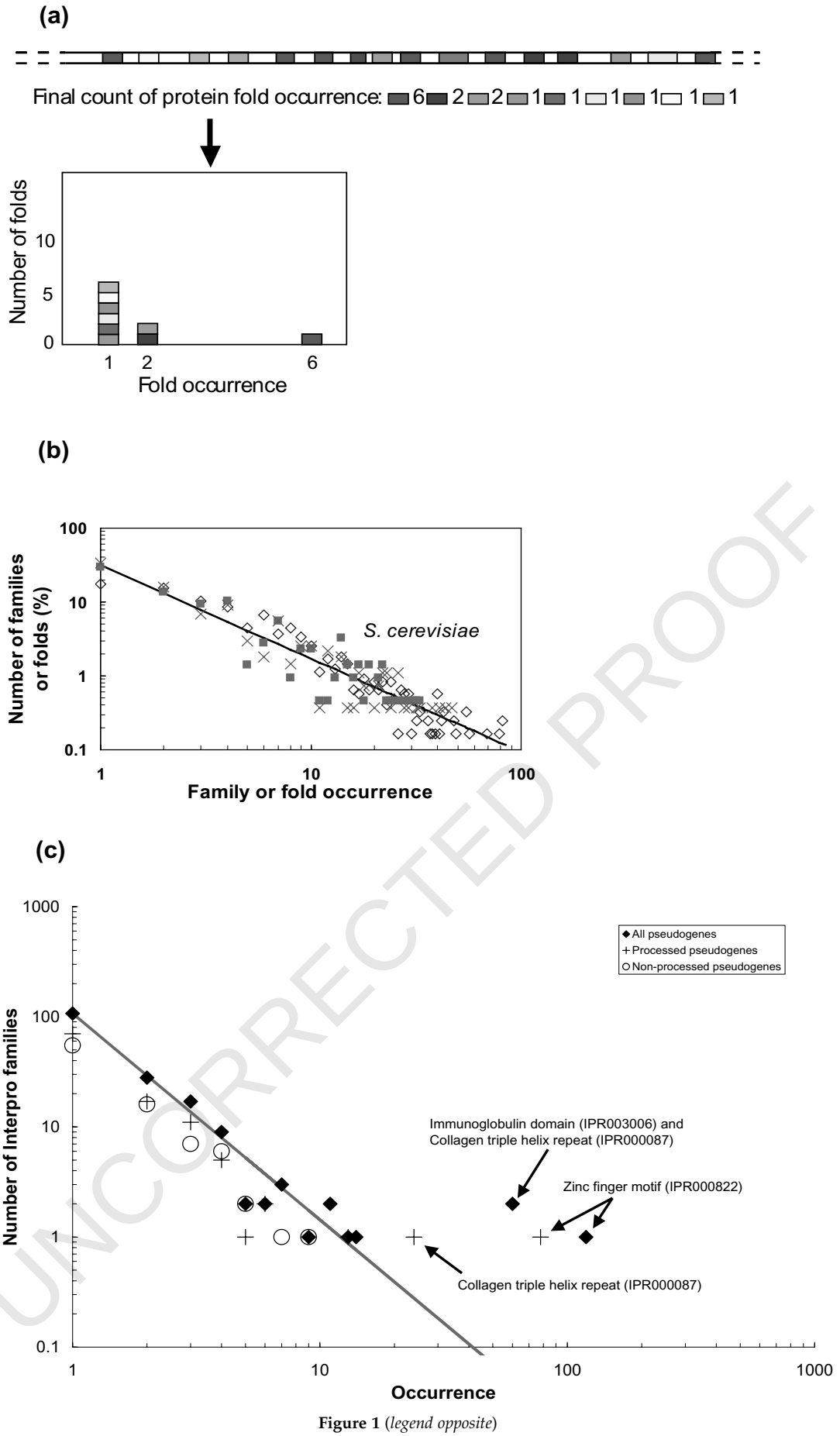

Figure 1 (*legend opposite*)

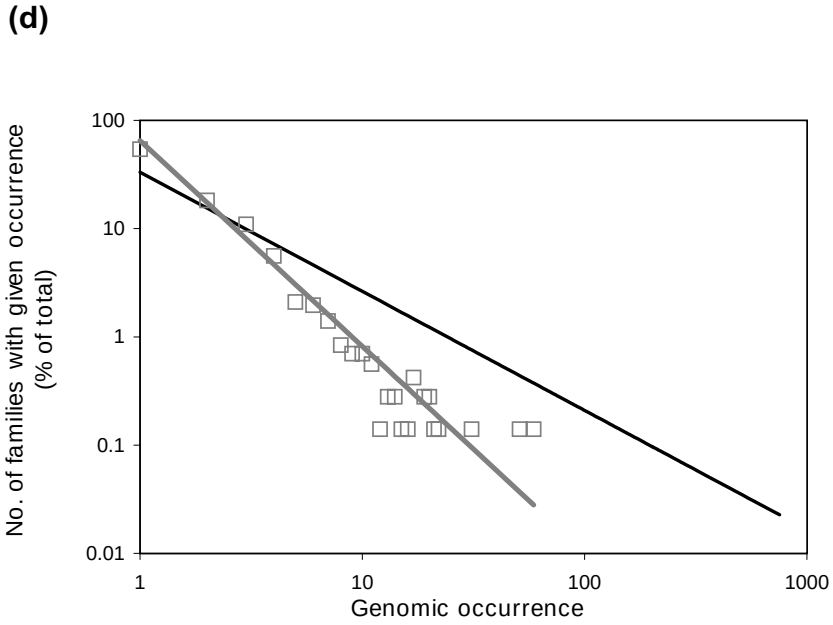

Figure 1. Power-law distribution of family sizes, superfamily sizes and protein fold occurrences in proteomes: adapted from Qian *et al.*,[22] (a) An illustration of how protein folds, superfamilies and families can be counted up, to give their total occurrences. (b) The power-law distribution of families (diamonds), superfamilies (crosses) and protein folds (filled squares) in the yeast (*S. cerevisiae*) proteome. The number of families or folds (*y*-axis) that have a particular occurrence (*x*-axis) is plotted. (c) Approximate power-law behaviour for InterPro protein sequence motifs in the pseudogene populations for human chromosomes 21 and 22 combined. The axes are as for (b). Outliers are labelled. (d) Power-law behaviour for a reliable subset of 1100 pseudogenes derived for the worm genome (see the text for details). The grey line is the power-law fit to the distribution for pseudogene families (open boxes); the black line is the same fit for the distribution for gene families, clustered as described.[69] The axes are as for (b).

comparable coverage of the SCOP domain database[8] in terms of superfamilies (between 460 (yeast) and 594 (human)[24]). Extensive sequence family redundancy is observed at the individual gene level in the eukaryotes, most notably in *Arabidopsis thaliana*, where only 35% of proteins are singletons (i.e. have no paralogs).[2] (For comparison, the degree of family redundancy is less extensive in the *Saccharomyces cerevisiae* genome, which by the same strict criteria, contains 29% of proteins in families.) In Arabidopsis, the extensive redundancy is linked to a large number of segmental chromosomal duplications arising from four distinct large-scale duplication events 100 to 200 million years ago.[29] Regardless of the mechanism of formation (whether segmental or local duplication), from an individual gene perspective, new gene duplicates in eukaryotes arise at the rate of about 0.01 per gene per million years, with rates for individual genomes ranging from 0.02 for *Caenorhabditis elegans* to 0.002 for *Drosophila melanogaster*; this is of the same order as the rate of mutation per nucleotide site.[30]

By what mechanism does the gene family redundancy chiefly arise? For example, Wolfe and colleagues identified homologous arrays of genes on different yeast chromosomes, which they hypothesized had arisen from a single, whole-genome duplication event about 100 million years ago, after separation from the *Saccharomyces kluyveri* yeast branch[31–33] However, ~90% of the resulting individual duplicated genes arising from this event appear to have been lost. Furthermore, there is no evidence that these duplications occurred at the same time; indeed, many segmental chromosomal duplications may have occurred in yeast at various times over the past 200–300 million years.[1] On the basis of the partial genome sequencing of 13 ascomycete relatives of *S. cerevisiae*, the conservation in yeast of singletons and gene family redundancy was found to arise mostly from local duplication events and did not support the whole-genome duplication hypothesis in yeast evolution.[34] Finally, in the human genome, notably, there is much less occurrence of pairs of chromosomal segments where the density of duplicated genes approaches that of *A. thaliana* or *S. cerevisiae*, indicating far less segmental chromosomal duplication.[27] Inclusion of detailed pseudogene annotations for the analysis described above would help to pin-point the mechanism of evolution of gene redundancy (see below for a discussion of pseudogene populations).

Indispensable and dispensable sub-proteomes

What is the minimal "indispensable" sub-proteome for the eukaryotic cell? Regardless of how the protein family redundancy in the yeast proteome has arisen, it seems clear from gene disruption experiments that the sub-proteome essential for yeast cell viability contains only ~1000 proteins.[35,36] This is about three times the number of proteins adjudged essential for a minimal prokaryotic cell.[37] Wagner noted, from analysis of gene disruption data for yeast, that there is no strong correlation between gene family redundancy and robustness against gene disruption. This indicates that there is a contribution to the robustness to mutation of a given gene that arises

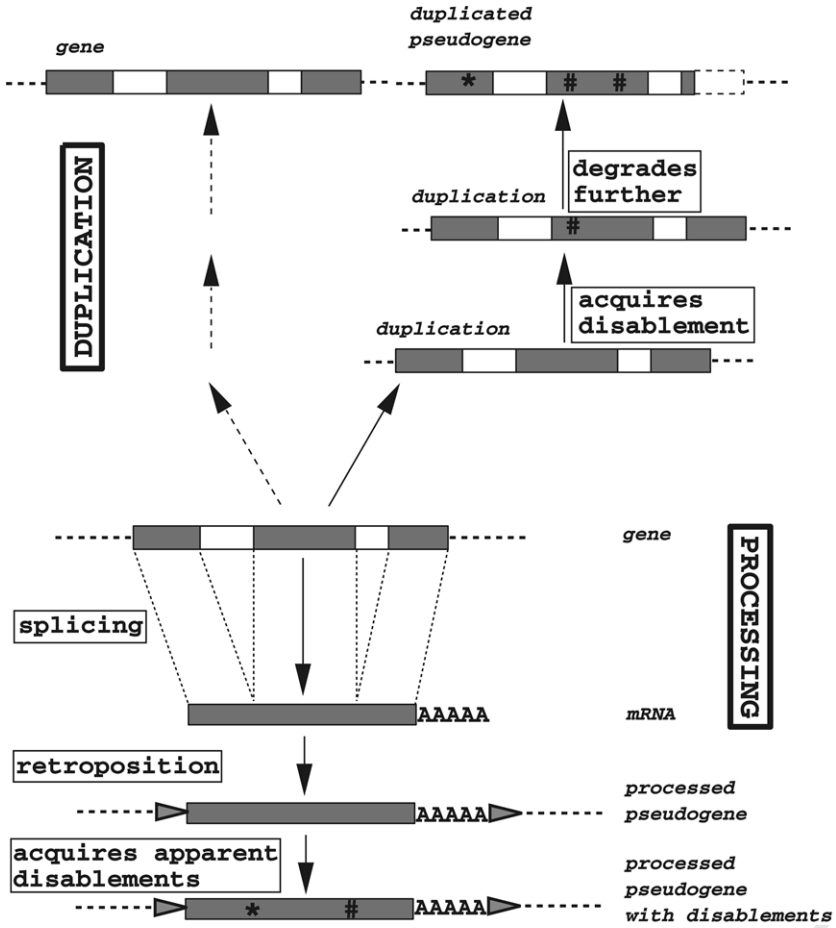

Figure 2. Two types of pseudogene. Pseudogenes are produced chiefly either by duplication or by processing. An example of a gene with three exons (shaded areas) is shown (boxed at the center of the Figure), with no non-coding segment in the exons for simplicity. ATG labels the start of the coding sequence, an asterisk (*) labels a stop codon and hash (#) stands for a frameshift. A non-processed or duplicated pseudogene simply arises when a gene duplication acquires a disablement that leads to: (i) lack of transcription; (ii) degradation *via* nonsense-mediated decay; or (iii) for an unknown subset of pseudogenes that produce messenger RNA transcripts escaping nonsense-mediated decay,[105] to degradation at some later unknown stage, so that a functioning protein chain is not formed. After such an initial disablement, the recently defunct pseudogene will acquire further obvious disablements of its reading frame (such as premature stops arising from point mutation, or truncations and frameshifts arising from deletion or insertion). A processed pseudogene arises when a messenger RNA transcript is reverse transcribed and re-integrated into the genomic DNA. Characteristic signals for these processed pseudogenes include small direct repeats (grey triangles) at either end of the pseudogene and a polyadenine tail (indicated here by AAAAA). The apparent coding frame of the pseudogene would then acquire obvious disablements, such as premature stops and frameshifts over evolutionary time.

from other genes with no detectable ancestral relationship, which, for instance, could provide alternative routes through pathways.[38]

From studies in yeast, it seems clear that many proteins have marginal effects on species fitness.[39] In a study of 34 *S. cerevisiae* genes that were judged non-essential by gene disruption,[35] 70% of them were found to have marginal but significant effects on the fitness of a strain.[40] This implies that the effective size of the indispensable sub-proteome for yeast can be determined only from study of its behaviour from generation to generation for the reproducing organism. This generation-weighted proteome could perhaps be dubbed the selectome, in analogy to the transcriptome (where the occurrence of different proteins is weighted by their transcription levels at different time-points and under various conditions[41,42]). The marginality of contribution to fitness in yeast, or protein dispensability, has been shown to be correlated with the molecular rate of evolution, i.e. more dispensable proteins evolve more rapidly.[43] It is conceivable that protein families with a higher molecular rate of evolution are more likely to have related pseudogenes in the genome. Proteins that have recently been dispensed with from the proteome may remain in the genome as pseudogenes (depending on genome-specific rates of genomic DNA loss and mutation), and this aspect of proteome evolution is discussed below.

The "dead" proteome: pseudogenes and proteome evolution

In the previous sections, we have discussed how the live part of the proteome of an organism is distributed into protein families, and some implications of this sequence redundancy. We now focus on the corresponding dead population of sequences, pseudogenes.

Pseudogenes are disabled copies of genes (or decayed remnants of genes) that do not produce a full-length protein chain. They can generally be divided into two types (Figure 2). Firstly, "processed" pseudogenes arise from reverse transcription from messenger RNA (mRNA) and re-integration into the genomic DNA.[44] These have

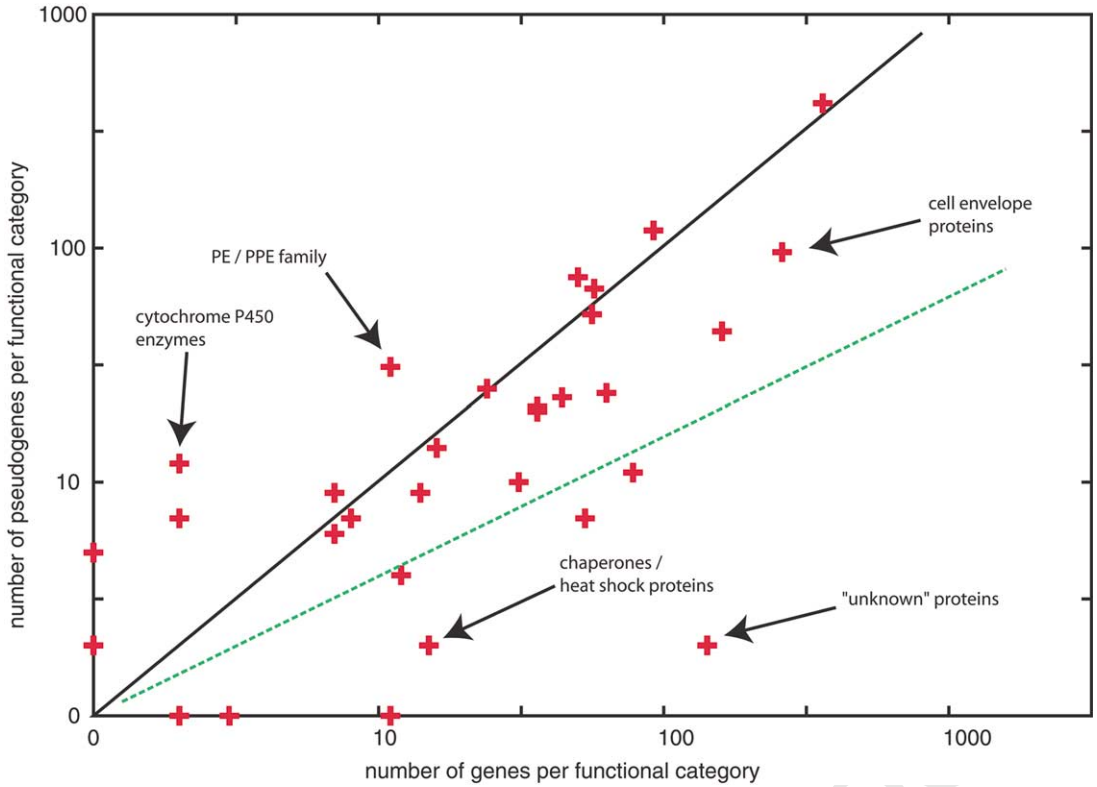

Figure 3. The relationship between the number of pseudogenes and genes for different functional categories in *M. leprae*. Each of the 31 functional categories listed by Cole *et al.*,[52] (Figure 2 of that paper) is plotted. The continuous line represents the number of pseudogenes being equal to the number of genes. Eleven of the categories are above this line, i.e. are more "dead" than "live". The dotted line represents the overall ratio of pseudogenes to genes in the proteome. Eight of the categories are below this line, i.e. more live than the overall ratio for live-to-dead.

been observed only in the metazoan animals and flowering plants, and presumably arise from mRNA transcripts in the germ-line cell lineage. In humans, they are probably made as a by-product of long interspersed nuclear element (LINE) retrotransposition.[45] After integration into the genome, they gradually accumulate disablements (stop codons, frameshifts, inserted repeat elements) of their reading frame. Secondly, "non-processed" or "duplicated" pseudogenes arise from duplication in the genomic DNA and subsequent disablement, most commonly through disruptive frameshift mutation or premature stop codon formation.[46] Formation of a pseudogene from gene duplication may have effects on the fitness of an organism; for example, if the duplicated gene has diverged very little since the duplication event that formed it (perhaps acquiring a slightly different activity or specificity in its function), the decrease in copy number for the gene family may be mildly deleterious. Conversely, copies of genes may be lost because that particular family is no longer as beneficial for fitness and has become more dispensable.

Pseudogenes, as "molecular fossils", are important sequences for the study of molecular evolution. Here, we discuss the occurrence of pseudogenes from a whole-proteome perspective, making use, where appropriate, of comparison of the prevalent families in proteomes and pseudogene populations. Such a perspective, of course, has been possible only recently with the advent of complete genome sequencing. We examine, in turn, the implications for proteome evolution in prokaryotes, and in the eukaryotes yeast, worm, fly and human. In prokaryotes, we see evidence for large-scale reductive evolution that mirrors the expansive evolution arising from horizontal transfer. In eukaryotes, we see that duplicated pseudogenes tend to be associated with environmental and response families. In the yeast, there appears to be a mechanism for conditionally "resurrecting" disabled genes as an evolutionary buffer to environmental fluctuation, perhaps in a concerted fashion. In the worm, the families of sequences that are prevalent in its pseudogene population have corresponding expanded or organism-specific populations in its genome, indicative of recent organism-specific expansions. For the fly, we argue that its apparently very small pseudogene population is linked to the size of its proteome through a very high rate of genomic DNA loss. Finally, for the human, we discuss the substantial number of processed pseudogenes relative to the putative total gene complement.

Prokaryotes: expansive and reductive proteome evolution

Prokaryotes can expand their proteomes by undergoing substantial horizontal transfer of genes from other strains and species.[47] Comparison of the two complete genomes sequences of *E. coli* strains O157:H7 EDL933 and K-12 MG1655,[48,49] shows how dramatically dynamic this horizontal transfer can be. Over a quarter (26%, 1387/5416) of the O157:H7 EDL933 genes are specific to that strain compared to K-12 MG1655. Conversely, in the same manner, 528/4405 (12%) of K-12 MG1655 genes are strain-specific. Strain-specific variation such as this has led some to argue that it is perhaps best to compare organisms in terms of a "species genome", with a core sub-proteome, and a variable sub-proteome that comprises the proteins and protein families that vary from strain to strain.[50,51] It will be interesting to see how closely correspondent such a core sub-proteome is to the indispensable subproteome, as discussed above for yeast.

Conversely, reductive evolution in bacteria may be equally dynamic. The recent sequencing of the genome of the bacterium *Mycobacterium leprae*, the leprosy pathogen, shows that it has undergone massive recent proteome decay[52] The *M. leprae* genome contains about ~1100 apparent pseudogenes, and ~1600 genes.[52] This is a considerable reduction when compared to the ~4000 proteins encoded in the genome of the related bacterium *Mycobacterium tuberculosis* and involves decrease in the redundancy of almost all protein families, with loss of substantial parts of pathways, such as the anaerobic respiratory chain. For example, the repetitive, glycine-rich PE and PPE families comprise 167 genes in the *M. tuberculosis* genome; however, in *M. leprae* it is more dead than live, there are only nine such genes in *M. leprae*, and 30 related pseudogenes. This family is shown on a plot for all of the functional classes reported here with pseudogene number plotted *versus* gene number (Figure 3). On the other hand, the functional class for chaperones and heat-shock proteins has a much smaller dead-to-live ratio than the overall ratio of dead to live proteins. By analogy with the two *E. coli* strains, it would be interesting to see to what extent the observed huge proteome decay is specific for the *M. leprae* strain sequenced, and how this affects the definition of its core sub-proteome.[50,51]

Proteome decay has been observed for two other pathogenic bacteria. The typhus pathogen *Rickettsia prowazekii* seems to have undergone such reductive evolution recently[53,54] Initially, it was thought to harbour only 12 pseudogenes,[53] but subsequently this estimate was enlarged. Prokaryote genomes are generally very compact, harbouring little non-coding genomic DNA (generally <10%; *E. coli* K-12 has ~11%[48]), implying that there is rapid deletion of any recently formed pseudogenes. However, the non-coding DNA in the *R. prowazekii* genome is >24% of the genomic DNA, suggesting that it comprises undetected decayed remnants of genes. Comparison of the *R. prowazekii* genomic sequence to those of other Rickettsias,[54,55] led to the detection of sequence similarity between (pseudo)genes in one species and the equivalent non-coding DNA in other species. These more fragmentary and disabled pseudogenic sequence homologies were dubbed fossil ORFs[55] or decayed orthologs.[54] Inclusion of these more decomposed remnants in *R. prowazekii* raises its total pseudogene population to 241 (compared to 834 live genes). The plague bacterium *Yersinia pestis* has a smaller relative proportion of pseudogenes (160, compared to ~4000 live genes) that appear linked to the loss of an enteropathogenic lifestyle[56]

Yeast: resurrectable variation between strains

There are very few annotated pseudogenes in the sequenced laboratory strain of *S. cerevisiae*, S288C;[57] we could find at most 30 such annotations in the SGD and MIPS databases.[11,58] From the analysis of disabled protein homology matches in the yeast genome, we believe that there may be up to a further 221 un-annotated pseudogenes in the *S. cerevisiae* S288C strain. This number rises further to 241 if we include pairs of existing ORF annotations, termed mORFs, that can be merged into a pseudogene and that could be complete ORFs in a different yeast strain[59] (Table 2). One of the most important previously documented pseudogenes in the yeast strain S288C is the FLO8 mutation.[60] This flocculin gene has an intact ORF in other strains but is disrupted by a single stop codon in S288C. This mutation has been shown to be the cause of the lack of diploid pseudohyphal filamentous growth in S288C, and has thus probably been selected in the laboratory so that yeast colonies are round and smooth. Strains that have an active FLO8 gene appear flocculent, having a fluffy colony appearance. The largest sequence families that are relatively prevalent in the S288C strain pseudogene population comprise flocculins like FLO8, the DUP family of double-transmembrane-helix proteins, growth inhibitors, helicases and stress-response proteins, whereas the most populated live families are forms of protein kinase, helicases, a transcriptional regulatory protein domain and the AAA ATPase domain (Table 3). Note how the pseudogenes appear to disproportionately have environmental and stress response functions. They have been found to occur near the ends of the chromosomes, mostly within 20 kb of the telomeres.[59]

Sup35p is part of the surveillance complex in yeast that controls translation termination and nonsense-codon read-through.[61,62] The $[PSI+]$ prion in yeast arises from the propagation of an alternatively folded amyloid-like form of Sup35p.[61,63] Thus, formation of the alternative

Table 2. Gene and pseudogene numbers

Organism	No. genes	No. pseudogenes	No. processed pseudogenes	No. duplicated pseudogenes	References
R. prowazekii (B)	834	241	0	241	53,54
M. leprae (B)	1604	1116	0	1116	52
Y. pestis (B)	4061	160	0	160	56
S. cerevisiae strain S288C (E)	6340	221 + 20 = 241[a]	0	241[a]	57,59
C. elegans (E)	20,009	1100 (2168)[b]	104 (208)	996 (1962)	66,69
D. melanogaster (E)	14,332	100 +	??	??	28; Harrison *et al.*, unpublished results
A. thaliana (E)	25,464	785	??	??	2
Homo sapiens (E)	~21,000 to ~39,000	??	~2900	??	27,85
Homo sapiens (E) (just chromosomes 21 + 22)	927	350	178	172	96

[a] This total is for dORFs plus mORFs. dORFs are pseudogenic or disabled ORFs that comprise a large fragment of disabled protein sequence homology that is not part of an existing ORF annotation; mORFs (merged ORFs) arise from merging two existing ORF annotations by ignoring their intervening stop codon.59

[b] This is for a set of disabled protein sequence homologies, supported by protein/cDNA/EST homology evidence. The values in parentheses are upper estimates derived as described.[69]

form of this protein takes Sup35p away from its normal functioning state, and can cause increased levels of nonsense-codon read-through in a particular strain, arguably leading to the full-length resurrection of ORFs that are apparently disabled. This can be seen as an evolutionary "buffering" effect, that enables a small amount of strain-specific variation to be maintained "in store". Indeed, the ability to form the [*PSI* +] prion itself may have been selected to enable this buffering effect. Interestingly, a recent study on [*PSI* +]-engendered phenotypic diversity, showed that one strain is more flocculent when in the [*PSI* +] state than in the [*psi* −] state;[64] this may be due to the resurrection of the complete FLO8 reading frame, or other flocculin genes.

Worm *versus* fly: comparison in terms of their live and dead proteomes

Despite their comparable genome size (100 Mb for the worm, 120 Mb euchromatic for the fly), and the greater apparent biological complexity of the fly (more cells, longer lifespan, more complicated physiology), the worm (at present) has more genes. The original sequencing projects estimated 19,099 worm and 13,601 fly proteins, although the proteomes comprise comparable functional diversity at the sequence domain level.[28,65–67] A recent gene prediction study for the fly genome has yielded 1042 additional candidate genes, potentially increasing the Drosophila gene total to >14,600 and the total proteome to >15,100.[68] Furthermore, alternative splicing for the fly may

Table 3. Comparison of the prevalent InterPro sequence motifs in the population of disabled ORFs and in the live proteome of yeast

Disabled ORFs/pseudogenes		Proteins	
No.	Description	No.	Description
12	WD40 (IPR001680)[a]	**115**	Eukaryotic kinase (IPR000719)
6	DUP membrane protein (IPR001142)	**112**	Serine–threonine protein kinase (IPR002290)
6	Mitochondrial electron transport (IPR001993)	**99**	WD40 (IPR001680)
6	Flocculin (IPR001389)	**76**	Dead-box helicase (IPR001410)
4	Helicase, C-terminal domain (IPR001650)	**74**	Helicase, C-terminal domain (IPR001650)
4	PIR repeat (IPR000420)	**57**	Fungal transcriptional regulatory protein (IPR001138)
3	BNR repeat (IPR002860)	57	AAA ATPase superfamily (IPR003593)
3	Zn-containing alcohol dehydrogenase (IPR002085)	**55**	TyA transposon protein (IPR001042)
3	Dead-box helicase (IPR001410)	**54**	RNA-binding region RNP-1 (IPR000504)
3	Fungal transcriptional regulatory protein (IPR001138)	53	C2H2 Zn finger (IPR000822)
3	SRP1/TIP1 stress-induced protein (IPR000992)		
3	DNA topoisomerase I DNA-binding domain (IPR003602)		

The name of each InterPro motif is given,[5] along with its number in parentheses. These counts are for the pseudogenic population derived from combining dORFs and mORFs (see the text and footnote **a** in Table 2).

[a] The 12 of these are all in one protein.

be more extensive than at present documented (currently about 2% of the documented worm proteome arises from alternative splicing, and ~7% for the fly).[28,65–67]

What about the corresponding sizes of the pseudogene populations for these two organisms? Depending on the thresholds used, the worm genome appears to contain a moderately sized complement of >1100 pseudogenes.[69] Only a small proportion (<~5%) of the pseudogenes appear to be processed. In general, the number of pseudogenes associated with each family of proteins is not proportional to the size of the family.[69] This would be the "default case" if duplicated pseudogenes were formed randomly from existing gene families. However, as shown in Table 4, the largest numbers of pseudogenes are associated with multiple families of seven-transmembrane chemoreceptors (these are also a class of "environmental response" proteins, which were observed above for yeast). Also common are families associated with a reverse transcriptase and a transposase, which presumably reflects remnants of decayed transposons (obvious transposons were screened out before the pseudogene assignment).

There are only 40 annotated pseudogenes for the fly genome, and a preliminary survey by the authors suggests at least ~60 more (P.M.H. *et al.*, unpublished results). (One should note, however, that an unknown number of gene annotations for either the fly or the worm may be shown to be pseudogenes, upon further characterization.) The cohort of olfactory receptors/chemoreceptors and other seven-transmembrane receptors in the worm (~1100) is almost a scale of magnitude larger than in the fly (~160 seven-transmembrane receptors). This perhaps indicates a recent evolutionary organism-specific expansion in these genes for the worm, or the converse (a contraction in number of members) for the fly.[65,66,70] Their predominance in the worm pseudogene population is presumably related to this apparent expansion of seven-transmembrane receptors in the worm. The substantial majority of these genes (~90%) appear to be organism-specific in the worm,[71] although careful sequence analysis using hidden Markov models has found mammalian orthologs for ~170 of them.[72] On a related note, of the estimated ~1000 seven-transmembrane olfactory receptor (pseudo)-genes in the human genome, about two-thirds are expected to be pseudogenic.[73,74]

Interestingly, the families that have the largest number of associated pseudogenes are amongst the families that are most expanded in the worm relative to the fly (Table 5). We compared in detail the list of domain sequence families for the fly and worm proteomes from the InterPro database.[5] The families exclusive in this list to either organism are tabulated, as well as the most expanded large families (with 30 or more members) relative to the other organism (Table 5). Three of the largest of these are for the seven-transmembrane receptor families (Table 5).

The small number of fly pseudogenes and the apparently small size of its proteome may be related to the overall genomic DNA deletion rate. The larger worm proteome may arise simply because factors such as genomic DNA deletion rates and chromosomal rearrangement have allowed it. It may be that the genomic DNA deletion rate in the fly (which was previously evidenced to be very high from the apparent rarity of true fly pseudogenes[75–77]) hampers the maintenance of recent gene duplications, so that they have less time to become evolutionarily useful. Experiments with transposable elements in *D. melanogaster* and the cricket genus Laupala indicate a very rapid loss of genomic DNA in Drosophila[78–80] Drosophila has an extremely high rate of chromosomal rearrangement.[81] However, studies on families of worm chemoreceptor genes and pseudogenes suggest that the worm has a rather high genomic DNA deletion rate.[70,82,83] Moreover, an analysis looking for small protein motifs selected from the Prosite database in intergenic regions in the fly and the worm suggests that the fly has as many, if not more, over-represented motifs (pseudomotifs) than the worm.[84] These pseudomotifs may represent fragments of protein fossils. Thus, their prevalence in the fly in relation to the worm, may indicate that the fly has much pseudogenic material that has decayed substantially.

Human: a large processed pseudogene population

For the human genome, the determination of the number of pseudogenes is intimately inter-linked with the determination of the total gene number, as cDNA/EST coverage for a full range of human tissues is likely to take many years. The recent near-complete sequencing of the human genome has yielded numbers for the human gene total that seem surprisingly low, of the order of 23,000–40,000 genes.[27,85] Efforts to estimate the number of human genes just prior to the publications of the sequenced genome, with one notable exception (which estimated ~120,000 human genes[86]), yielded largely similar numbers to these, in the range ~28,000 to ~35,500.[87–90] A recent comprehensive annotation of the draft human genome estimated about 65,000–75,000 transcriptional units or genes in the genome.[91]

Duplicated pseudogenes are more involved in the problem of gene prediction than processed pseudogenes: an exon with a disablement that is in the region of a gene may or may not be a part of the extant gene, making it difficult or impossible to determine if the gene is a pseudogene without cDNA/EST evidence. This is compounded by the prevalence of alternative splicing in the human genome; three independent surveys have shown that ~40% of genes encode alternatively spliced transcripts.[92–94] Estimates for the proportion of gene annotations

Table 4. Largest families in terms of proteins and pseudogenes in the worm; adapted from previous family clustering[69]

Pseudogenes		Proteins	
No.	Description	No.	Description
59	Reverse transcriptase (IPR000477)	**216**	Nuc. hormone receptor ligand-binding domain (IPR000536)
51	7-TM chemoreceptor family #1 (IPR000168, IPR003003)	**193**	7-TM chemoreceptor family #1 (IPR000168, IPR003003)
31	Unknown domain family #1[a]	**188**	7-TM chemoreceptor family #2 (IPR000168)
27	7-TM chemoreceptor family #2 (IPR000168)	**124**	Eukaryotic kinase (IPR000719)
22	7-TM chemoreptor family #3 (IPR000168)	**93**	MATH domain (IPR002083)
21	Major sperm protein (IPR000535)	**70**	7-TM receptor family #4 (IPR000276)
20	Unknown domain family #3[a]	**70**	Guanylyl cyclase recep. tyr kinase (IPR001054)
19	Unknown domain family #4[a]	**70**	Cytochrome P450 (IPR001128)
19	TcA transposase (IPR002492)	**70**	Tyr phosphatase (IPR000242)
17	7-TM receptor family #4 (IPR000276)	**68**	UDP-glucuronyl transferase (IPR002213)

Corresponding InterPro motifs for some families are indicated in brackets. The thickly outlined boxes are for families that occur in both the top ten pseudogenes and top ten protein families.
[a] Those families do not have corresponding InterPro motifs.

Table 5. Exclusive and expanded large families for assigned INTERPRO domains in the fly and worm proteomes

Largest exclusive to fly[a]		Largest exclusive to worm[b]		Most expanded in worm relative to fly[b]		Most expanded in fly relative to worm	
No.	Description	No.	Description	No. in worm (fly)	Description	No. in fly (worm)	Description
404	Insect cuticle protein (IPR000618)	**624**	7-TM chemo-receptor family (IPR000168, IPR003003)	**60 (1)**	DUF23 (IPR002875)	**544 (15)**	Chymotrypsin serine protease family S1 (IPR001314)
110	Alkaline phosphatase (IPR001952)	**322**	7-TM chemo-receptor family (IPR000168)	**301 (6)**	EB module (IPR002899)	**950 (35)**	Serine protease trypsin family (IPR001254)
99	Glycoside hydrolase family 22 (IPR001916)	**276**	DUF38 (IPR002900)	**44 (1)**	ET module (IPR002603)	**161 (6)**	Lipase (IPR000734)
73	Alpha-tocopherol transport protein (IPR001071)	**238**	ShK toxin domain (IPR003582)	**339 (8)**	MATH domain (IPR002083)	**48 (2)**	Peptidyl di-peptidase A M2 metallo-protease (IPR001548)
54	Hemocyanin (IPR000896)	**237**	DUF139 (IPR003341)	**58 (2)**	K + channel (IPR003280)	**38 (2)**	GMC oxido-reductase (IPR000172)
30	Acylphosphatase (IPR001792)	**233**	7-TM chemo-receptor family (IPR000168)	**167 (6)**	Major sperm protein (IPR000535)	**37 (2)**	NMDA receptor (IPR001508)
29	GYR motif (IPR004011)	**184**	pol-like reverse transcriptase (IPR003286)	**71 (3)**	TcA transposase family (IPR002492)	**44 (3)**	Chaperonin cpn60 60 kDa sub-unit (IPR001844)
26	Mitochondrial brown fat uncoupling protein (IPR002030)	**148**	SRG family integral membrane protein (IPR000609)	**438 (37)**	Nuclear hormone receptor ligand-binding domain (IPR000536)	**47 (4)**	Gamma tubulin (IPR002454)
25	Opsin (IPR001760)	**145**	Nematode cuticle collagen N-terminal domain (IPR002486)	**861 (75)**	F box domain (IPR001810)	**35 (3)**	Neutrophil cytosol factor 2 (IPR000108)
25	NF-κB/Rel/dorsal (IPR000451)	**109**	WSN (domain of unknown function) (IPR003125)	**167 (17)**	vWF type A domain (IPR002035)	**76 (9)**	Insect alcohol dehydrogenase (IPR002424)

These data are taken from the lists provided on the InterPro proteome analysis Website (http://www.ebi.ac.uk/proteome). The symbols and abbreviations are explained in Table 4. The boxed families occur also in the top ten pseudogene family list for worm.

[a] The four lists are sorted in decreasing order of the degree of expansion. The degree of expansion in a family is simply the size of the family in one organism divided by its size in the other. Only families with 30 or more members in either organism are considered for this analysis.

[b] The family numberings differ here from those in Table 2, as these are derived by motif scanning in individual sequences, whereas the Table 2 families are derived by our own sequence clustering procedure (see Table 2).

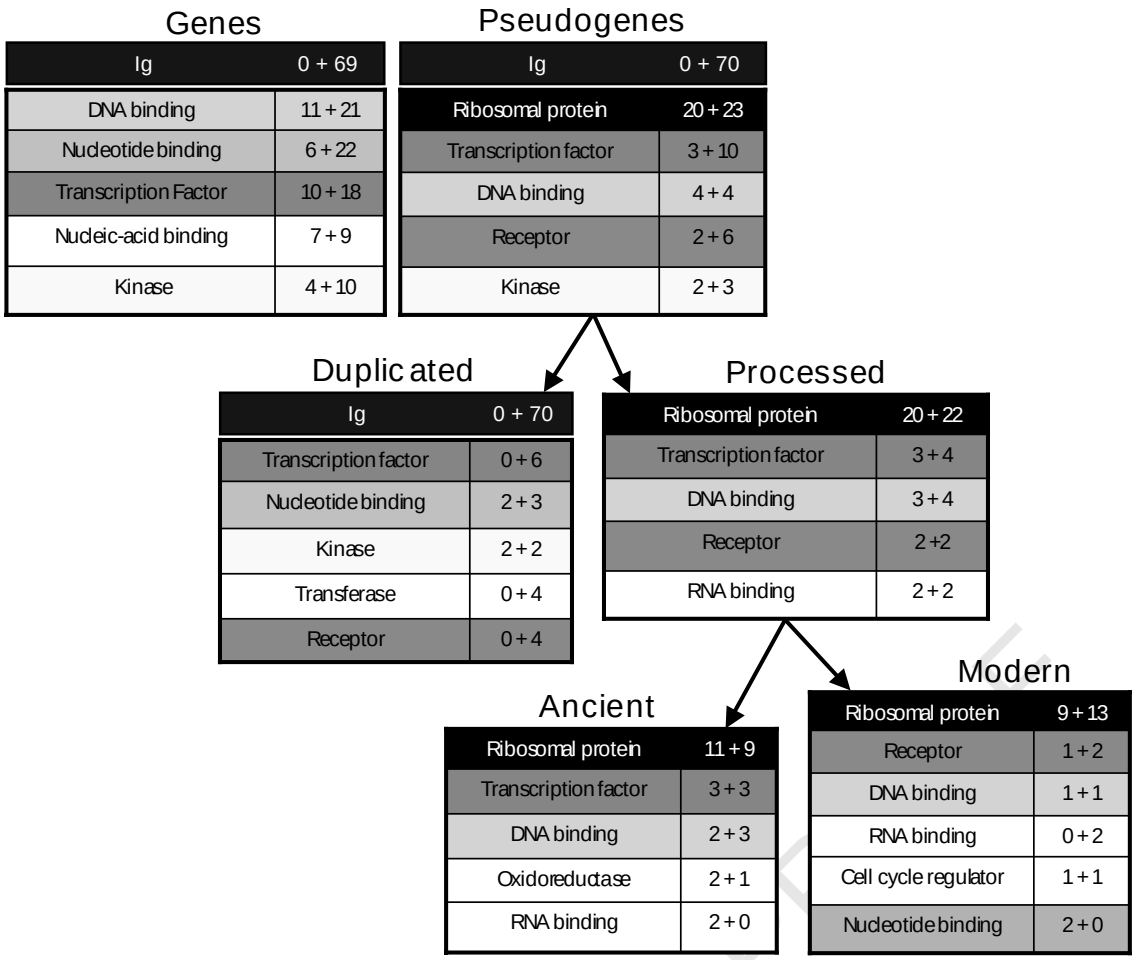

Figure 4. Functional categories of genes and pseudogenes in chromosomes 21 and 22: adapted from data given by Harrison *et al.*,[96] Gene Ontology (GO) functional classes were assigned to predicted genes and pseudogenes for chromosomes 21 and 22 in combination. Those for pseudogenes are separated into processed and duplicated, with processed pseudogenes further separated into ancient and modern processed pseudogenes on the basis of their degree of sequence identity with the closest-matching human gene from the Ensembl data set (http://www.ensembl.org

that may actually be pseudogenes lie in the range 4–22%.[27,87,95]

Processed pseudogenes will be less likely to interfere with the accuracy of gene predictions; they will, on average, tend to be longer than the average human exon size, and comprise characteristic signals, including a C-terminal polyadenine tail.[44,46] If they occur in relatively large numbers, they are also, in a sense, evidence that their parent gene is transcribed and most likely functional. Estimated numbers of processed pseudogenes in the human genome are substantial compared to those estimated for the gene total. In the completed chromosome 22 sequence, Dunham *et al.* initially predicted at least 545 genes and 134 pseudogenes (one for every ~4.1 genes).[87] They surmised that 82% of these pseudogenes were processed, as they contained single blocks of homology and lacked the characteristic exonic structure of the closest matching gene. This gives a predicted proportion of one processed pseudogene for every ~5.0 genes. Venter *et al.*, observed evidence for at least ~2900 processed pseudogenes arising from their human gene set.[85] These were identified by searching for continuous spans of homology of >70% sequence identity over >70% of the length of the matching coding sequences from their gene annotations. No effort was made to look for the other characteristics of processed pseudogenes, such as evidence for polyadenylation. This data set of processed pseudogenes gives a smaller proportion of processed pseudogenes, in the region of about one for every ten genes. A survey by the authors of pseudogenes on chromosomes 21 and 22 that included searching for polyadenylation yielded an estimate of about one processed pseudogene for every four genes.[96] In this survey, we found that about half of all detected pseudogenes are processed (Table 2). The large amount of processing in the human genome may simply reflect its large amount of intergenic sequence and perhaps, the genomic mobility of transposable elements such as LINE-1.[45]

The prevalence of the encoded proteins in the processed pseudogene population appears to be related to expression. Goncalves *et al.* analysed 181 genes that were reported to have one or more processed pseudogenes.[97] They found that such genes tend to be short, highly conserved and widely expressed. In the survey of ~2900 potential processed pseudogenes by Venter *et al.*[85] (noted above), by far the most prevalent class of transcripts (>60%) were for ribosomal proteins, which are very highly (and, of course, widely) expressed. The possibility of a large number of processed pseudogenes for ribosomal proteins was first noted during cloning of the mouse ribosomal protein rpL32[98] As shown in Figure 4, data by the authors from a survey of chromosomes 21 and 22 for processed and duplicated pseudogenes[96] also indicate that ribosomal proteins predominate in the processed pseudogene population, albeit, to less of an extent than in the survey by Venter *et al.*,[85] we found that ~20% of processed pseudogenes were ribosomal, and that there was little difference in this prevalence for either modern or ancient processed pseudogenes.

Figure 4 shows that the duplicated pseudogenes found in the survey of chromosomes 21 and 22 tend to be immunoglobulin gene fragments, reflecting their prevalence on chromosome 22. This preference continues the environmental-response theme discussed above for the worm and the yeast.

Concluding remarks

Comparing and contrasting the distribution of protein families in proteomes and in pseudogene populations gives us new perspectives on how proteomes evolve. A number of over-arching themes and implications are apparent. There are three distinct populations of pseudogenes.

Three types of pseudogenes

Prokaryotic pseudogene: dying genes resulting from a niche change

Prokaryotic pseudogenes appear to be genes that are dying and disappearing from the genome, in response to a fundamental niche change for an organism. In particular, there are now three bacterial pathogenic genomes (*M. leprae*, *Y. pestis* and *R. prowazekii*) that exhibit large-scale degradation of the proteome, with the lost or depleted families evidencing apparent niche change. In the most extreme case, *M. leprae* has large-scale patterning in its pseudogene population that indicates modular loss of metabolic pathways and branches of pathways, such as part of the anaerobic respiratory chain, when compared with *M. tuberculosis*, its closest sequenced relative. It is interesting, however, that this organism has lost *dnaQ*-mediated proofreading activities of DNA polymerase III.[52] Perhaps, this loss of function may actually have been selected so that removal of redundant genes could be accelerated. Although selection for deletion of pseudogenic DNA may not be sufficiently strong in eukaryote genomes,[79] there may be strong selection pressures for such deletion in small prokaryotic genomes that are undergoing niche change, and discarding many genes.

Eukaryotic processed pseudogenes: random insertion events

Processed pseudogenes arise from reverse-transcription of mRNA and re-integration into the genome. In humans, they are probably made as a by-product of LINE retrotransposition[45,99] That is, the processed pseudogene is formed from reverse transcribing a spliced mRNA into a cDNA using the reverse transcriptase from the LINE and re-integrating into the genome.[45,99] Initial surveys suggest that their occurrence is largely based on simply random insertions, with their prevalence based on (1) the amount of mRNA to be inserted (expression levels) and (2) the amount of intergenic DNA available for insertion. The first factor accounts for the large numbers of ribosomal protein families found in processed pseudogenes.[85,96] The second factor explains the large number of processed pseudogenes in the human genome, relative to the worm. It appears that the number of processed pseudogenes per 10^6 bases of non-coding DNA is almost the same for both organisms. For human (chromosomes 21 and 22) the ratio is 2.6, which is 178 processed pseudogenes per 67 Mb of non-coding DNA. For the worm, the comparable number is 3.0, which is 208 per 70 Mb. (This ratio uses the high estimate for numbers of pseudogenes in the worm. It would decrease by 50% if one used the lower estimate (see Table 2).)

Eukaryotic duplicated pseudogenes: a resurrectable reservoir of extra parts for environmental response?

Duplicated eukaryotic pseudogenes appear to be most intriguing. They tend to arise for organism-specific environmental response functions. This tendency may reflect genomic mechanisms that an organism uses to generate proteins that deal with changes in its environment. We suggest below that pseudogenes or pseudogenic parts for such classes of gene may occasionally be resurrected and used to enable larger random leaps in sequence space (see below).

Eukaryotic pseudogenes tend to occur for organism-specific families. Pseudogenes in yeast are about twice as likely as a live protein to be yeast-specific.[59] Similarly, in the worm, the vast majority of the most prominent pseudogene families (those for the 7-TM chemoreceptors, major sperm protein and some unknown domains) are worm-specific or represent families vastly

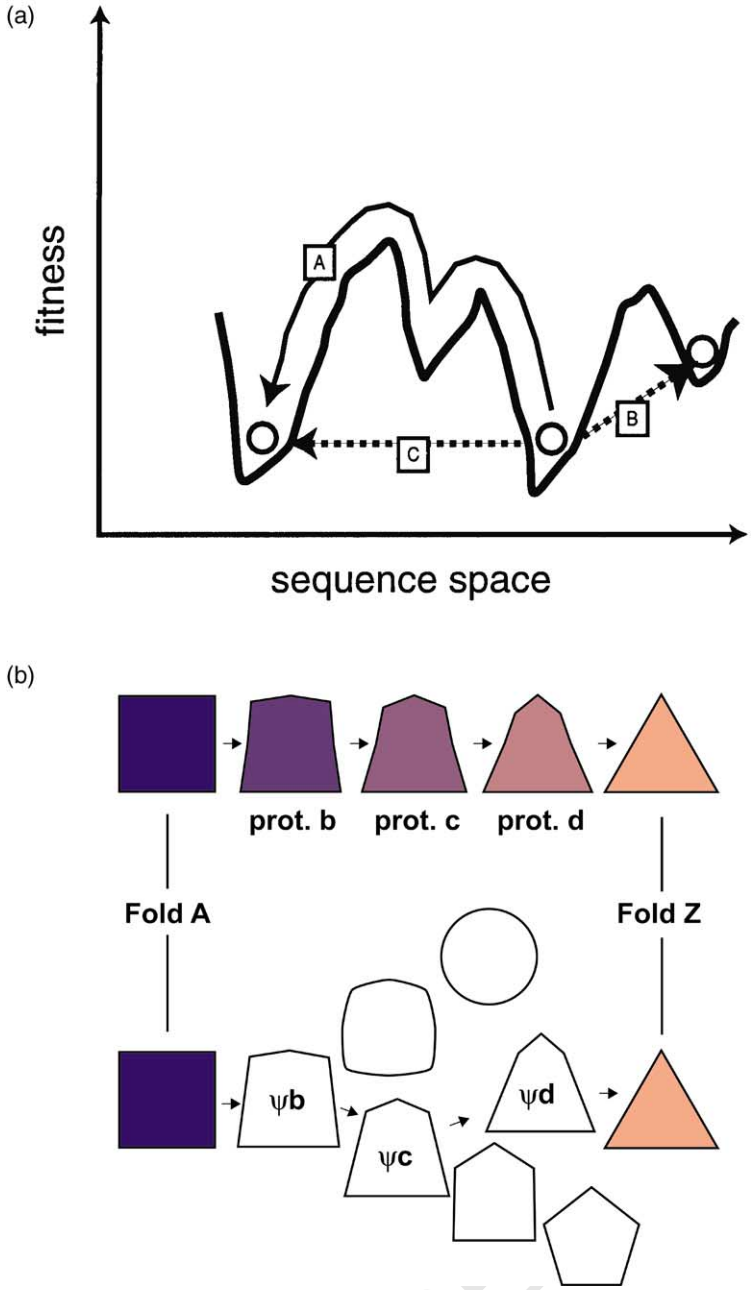

Figure 5. Aspects of pseudogene resurrection as an evolutionary mechanism. (a) A schematic evolutionary landscape showing a sequence (represented by an open circle) in a favourable fitness minimum, with three evolutionary routes A, B and C. Route A (continuous line) arises from mutation under the pressures of natural selection. Route B (dotted line) represents what happens when a sequence undergoes random drift as a pseudogene, but which, when "resurrected" as a genic sequence, is unfit. Route C represents what happens when a sequence undergoes random drift as a pseudogene, but reaches another favourable fitness minimum in a shorter span of time than would be possible under continuous natural selection. (b) The top panel shows the conventional view of protein fold evolution where every intermediate along the pathway has to be transcribed and translated. The bottom panels shows a pathway that involves pseudogenic fragments.

expanded in the worm relative to the fly (Tables 4 and 5).

Pseudogenicity in eukaryotes appears to be linked to protein functions that are needed for environmental response, needing functional "breadth". In the worm, pseudogenicity is linked to 7-TM chemoreceptor families.[69] In the yeast, flocculins (which perform a variety of functions involving cell adhesion), growth-inhibitors, and stress-response proteins have the highest numbers of pseudogenes.[59] Finally, in the human, immunoglobulins have a high degree of pseudogenicity. For example, the immunoglobulin locus containing lambda variable-region gene segments on chromosome 22 is about 50% pseudogenic.[96] Also, a recent survey shows that there are ~1000 olfactory receptors in the human genome, with 60% of these pseudogenic.[73]

Pseudogene resurrection as a general evolutionary mechanism

In certain cases, as a rare or occasional evolutionary event, the resurrection of duplicated pseudogenic DNA to an expressed protein may enable sampling of more sequence space for a protein or protein family (Figure 5(a)). In particular, pseudogenes or parts of pseudogenes may be re-used, after having drifted randomly without selection for a period of evolution. The idea of such "untranslatable intermediates" in the evolution of a protein was first postulated about 30 years ago

by Koch.[100] Although generally one would expect this mechanism to produce unviable or unfavourable leaps in sequence space, occasionally it may provide a shorter evolutionary route to another favourable evolutionary energetic minimum (Figure 5(a)).

There are number of cases that one can point to as evidence of such resurrection. A pseudogene of bovine seminal ribonuclease that lay dormant for ~20 million years, appears to have been resurrected to form a functioning gene, probably *via* a gene conversion event.[101] As discussed above, the presence of the [PSI +] prion in yeast strains may enable resurrection or extension of ORFs from the yeast genome that have been able to drift without selection pressures since the occurrence of their disrupting mutations.[64] The large cohort of pseudogenes for chemo- or olfactory receptors (ORs) in metazoans (60% of the ORs in the human genome are pseudogenic) may be resurrectable by gene conversion events. There appears to have been a large number of gene conversion events (>20) in a cluster of olfactory receptors on chromosome 17 over the course of primate evolution.[102] This cluster contains 16 OR genes and 6 OR pseudogenes in the human genomic DNA. Gene conversion events in OR gene clusters may help to generate diversity at the odorant binding site.[102] Occasional resurrection of OR pseudogenes by gene conversion may contribute to this generation of diversity in binding capability. Finally, in the chicken, diversity of immunoglobulin heavy chain variable-region gene segments appears to be generated by gene conversion of a single functional gene with >80 pseudogenic gene segments.[103]

Resurrectable pseudogenes may help resolve a paradox about protein fold evolution

Considering duplicated pseudogenes as a resurrectable reservoir of diversity may help to resolve an evolutionary paradox presented by structural biology. How do new folds evolve? An early observation from structural genomics analyses was that there appear to be folds unique to certain phylogenetic groups.[16,25] For instance, an initial analysis showed that of 275 folds, 46 were present only in eubacteria and 73 only in eukaryotes, and of the 229 total folds in eukaryotes, 20 were only in plants and 90 only in animals.[16] How does one get new unique folds in certain phylogenetic groups? As shown in Figure 5(b), in some cases it may be difficult to imagine a scenario for this where each intermediate form has to be a functioning protein that is transcribed and translated. (This is in contrast to other evolutionary pathways, where functioning and selected intermediates are more plausible.) One can speculate that resurrectable pseudogenes could eliminate this paradox to some degree. A sequence comprising a particular domain fold or (more likely) part of a domain could become pseudogenic. It could then drift freely as a pseudogene, and evolve to a new domain fold upon or after resurrection. In this scheme, each intermediate does not have the constraint that it be a folded functional protein.

Elimination of pseudogenes

Pseudogenes can be eliminated from the genome due to deletion events. There is obviously greater pressure to do this for prokaryotes than for eukaryotes. Thus, it is important to point out that the lack of a large pseudogene population for prokaryotes does not imply that an organism has not undergone gene loss as drastic as that seen in *M. leprae*, over a similar evolutionary period. An organism with a higher rate of genomic DNA deletion would delete pseudogenic DNA more efficiently, and we would therefore not see such a large pseudogene population at present. For *M. leprae*, it may be that the rate of disablement of ORFs is raised, without there being a concomitant increase in the rate of deletion of intergenic DNA. Rates of intergenic DNA deletion vary widely from organism to organism.[80] For the eukaryote Drosophila, although the overall genomic deletion rate is very high, the observed spectrum of deletion sizes in transposable elements implies that it has not been selected for to aid genome compaction.[79] The Drosophila genomic DNA deletion rate seems to explain the dearth of pseudogenes in the fly that are detectable by sequence homology.[78,80] To find very decayed remnants of proteins in the genome not amenable to sequence alignment, we are currently developing a probabilistic approach based on scanning the genome for decayed protein motifs (termed pseudomotifs).[84] Over even longer evolutionary periods, gene loss can be inferred from careful comparative proteome analysis. For example, comparison of the *S. cerevisiae* proteome with the near-complete proteome of the fission yeast *Schizosaccharomyces pombe*, indicates the possible loss of about 300 proteins in *S. cerevisiae*, and provides an explanation for the small degree of gene splicing in *S. cerevisiae*, involving deletion of signalosome and spliceosome components.[104] (The fission yeast has extensive gene splicing.)

Power-law behaviour and the size of duplicated pseudogene populations

We noted above that the size of protein families in the live proteomes is governed by a power-law distribution (Figure 1). This behaviour is observed for the distribution of protein families in the pseudogene population (the dead proteome) of chromosomes 21 and 22, and of the worm genome[69,96] (Figure 1). (It is observed even for the distribution of pseudomotifs in the fly and worm genomes.[84]) This may imply that conservation pressures do not cause such power-law behaviour, but rather the flow of change from old to new families over evolution. Qian *et al.*,[22] found that the power-law distribution of protein families and

folds is well described by a simple model in which existing gene sequences can be duplicated, but with the occasional creation or addition of a novel gene.

Thus, despite the great differences in specific protein families prevalent in various organisms in both the living and the dead proteomes, we can see a clear commonality in their occurrence: one has a few families occurring many times and most occurring just a few times. In all aspects of genomic biology, one never gets a uniform distribution of occurrence over families.

Acknowledgments

Thanks to Julian Gough (MRC) for providing data on protein folds in eukaryotic proteomes, and to Nick Luscombe and Jiang Qian for part of Figure 1. Thanks also to Hedi Hegyi, Suganthi Balasubramaniam, Paul Bertone, Nathaniel Echols, Nick Luscombe and Ted Johnson for help, and to Alan Weiner for comments on pseudogene formation. M.G. acknowledges support from the Keck foundation and the NIH structural genomics initiative (P50 GM62413-01).

References

1. Friedman, R. & Hughes, A. L. (2001). Gene duplication and the structure of eukaryotic genomes. *Genome*, **11**, 373–381.
2. Arabidopsis Genome Initiative, T. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
3. Huynen, M. A. & van Nimwegen, E. (1998). The frequency distribution of gene family sizes in complete genomes. *Mol. Biol. Evol.* **15**, 583–589.
4. Nevill-Manning, C. G., Wu, T. D. & Brutlag, D. L. (1998). Highly specific protein sequence motifs for genome analysis. *Proc. Natl Acad. Sci. USA*, **95**, 5865–5871.
5. Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M. *et al.* (2000). InterPro—an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics*, **16**(12), 1145–1150.
6. Yona, G., Linial, N. & Linial, M. (2000). ProtoMap: automatic classification of protein sequences and hierarchy of protein families. *Nucl. Acids Res.* **28**(1), 49–55.
7. Krause, A., Stoye, J. & Vingron, M. (2000). The SYSTERS protein sequence cluster set. *Nucl. Acids Res.* **28**(1), 270–272.
8. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**(4), 536–540.
9. Pearl, F., Todd, A. E., Bray, J. E., Martin, A. C., Salamov, A. A., Suwa, M. *et al.* (2000). Using the CATH domain database to assign structures and functions to the genome sequences. *Biochem. Soc. Trans.* **28**(2), 269–275.
10. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M. *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25**(1), 25–29.
11. Mewes, H. W., Frishman, D., Gruber, C., Geier, B., Haase, D., Kaps, A. *et al.* (2000). MIPS: a database for genomes and protein sequences. *Nucl. Acids Res.* **28**(1), 37–40.
12. Riley, M. & Space, D. B. (1996). Genes and proteins of *Escherichia coli* (GenProtEc). *Nucl. Acids Res.* **24**(1), 40.
13. Gerstein, M. & Hegyi, H. (1998). Comparing genomes in terms of protein structure: surveys of a finite parts list. *FEMS Microbiol. Rev.* **24**, 1–28.
14. Gerstein, M. (1998). How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Fold. Des.* **3**, 497–512.
15. 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. *Nucl. Acids Res.* **25**(17), 3389–3402.
16. Gerstein, M. & Levitt, M. (1997). A structural census of the current population of protein sequences. *Proc. Natl Acad. Sci. USA*, **94**(22), 11911–11916.
17. Sonnhammer, E. L. & Durbin, R. (1997). Analysis of protein domain families in *Caenorhabditis elegans*. *Genomics*, **46**(2), 200–216.
18. Salamov, A. A., Suwa, M., Orengo, C. A. & Swindells, M. B. (1999). Genome analysis: assigning protein coding regions to three-dimensional structures. *Protein Sci.* **8**(4), 771–777.
19. Hegyi, H., Lin, J. & Gerstein, M. (2002). Structural genomics analysis for twenty proteomes: common, shared and unique protein folds. *Proteins: Struct. Funct. Genet.* in press.
20. Wolf, Y. I., Brenner, S. E., Bash, P. A. & Koonin, E. V. (1999). Distribution of protein folds in the three superkingdoms of life. *Genome Res.* **9**, 17–26.
21. Wolf, Y. I., Grishin, N. V. & Koonin, E. V. (2000). Estimating the number of protein folds and families from complete genome data. *J. Mol. Biol.* **299**, 897–905.
22. Qian, J., Luscombe, N. M. & Gerstein, M. (2001). Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *J. Mol. Biol.* **313**(4), 673–681.
23. Lin, J. & Gerstein, M. (2000). Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. *Genome Res.* **10**, 808–818.
24. Gough, J. & Chothia, C. (2002). SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucl. Acids Res.* **30**(1), 268–272.
25. Gerstein, M. (1997). A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J. Mol. Biol.* **274**(4), 562–576.
26. Yanai, I., Camacho, C. J. & DeLisi, C. (2000). Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification. *Phys. Rev. Letters*, **85**, 2641–2644.
27. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J. *et al.* (2001). Initial sequencing and analysis of the human genome. International Human Genome Sequencing Consortium. *Nature*, **409**(6822), 860–921.
28. Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G. *et al.*

(2000). The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.
29. Vision, T. J., Brown, D. G. & Tanksley, S. D. (2000). The origins of genomic duplications in Arabidopsis. *Science*, **290**, 2114–2117.
30. Lynch, M. & Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, **290**(5494), 1151–1155.
31. Seoighe, C. & Wolfe, K. H. (1999). Yeast genome evolution in the post-genome era. *Curr. Opin. Microbiol.* **2**(5), 548–554.
32. Seoighe, C. & Wolfe, K. H. (1999). Updated map of duplicated regions in the yeast genome. *Gene*, **238**(1), 253–261.
33. Wolfe, K. H. & Shields, D. C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, **387**(6634), 708–713.
34. Llorente, B., Durrens, P., Malpertuy, A., Aigle, M., Artiguenave, F., Blandin, G. *et al.* (2000). Genomic exploration of the hemiascomycetous yeasts: 20. Evolution of gene redundancy compared to *Saccharomyces cerevisiae*. *FEBS Letters*, **487**(1), 122–133.
35. Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K., Andre, B. *et al.* (1999). Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science*, **285**, 901–906.
36. Delneri, D., Brancia, F. L. & Oliver, S. G. (2001). Towards a truly integrative biology through the functional genomics of yeast. *Curr. Opin. Biotech.* **12**, 87–91.
37. Mushegian, A. (1999). The minimal genome concept. *Curr. Opin. Genet. Dev.* **9**(6), 709–714.
38. Wagner, A. (2000). Robustness against mutations in genetic networks of yeast. *Nature Genet.* **24**, 355–361.
39. Tautz, D. (2000). A genetic uncertainty problem. *Trends Genet.* **16**, 475–477.
40. Thatcher, J. W., Shaw, J. M. & Dickinson, W. J. (1998). Marginal fitness contributions of nonessential genes in yeast. *Proc. Natl Acad. Sci. USA*, **95**, 253–257.
41. Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E. *et al.* (1997). Characterization of the yeast transcriptome. *Cell*, **88**(2), 243–251.
42. Jansen, R. & Gerstein, M. (2000). Analysis of the yeast transcriptome with structural and functional categories: characterizing highly expressed proteins. *Nucl. Acids Res.* **28**(6), 1481–1488.
43. Hirsh, A. E. & Fraser, H. B. (2001). Protein dispensability and rate of evolution. *Nature*, **411**, 1046–1049.
44. Vanin, E. F. (1985). Processed pseudogenes: characteristics and evolution. *Annu. Rev. Genet.* **19**, 253–272.
45. Esnault, C., Maestre, J. & Heidmann, T. (2000). Human LINE retrotransposons generate processed pseudogenes. *Nature Genet.* **24**, 363–367.
46. Mighell, A. J., Smith, N. R., Robinson, P. A. & Markham, A. F. (2000). Vertebrate pseudogenes. *FEBS Letters*, **468**, 109–114.
47. Eisen, J. A. (2000). Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. *Curr. Opin. Genet. Dev.* **10**, 606–611.
48. Blattner, F. R., Plunkett, G., III, Bloch, C. A., Perna, N. T., Burland, V., Riley, M. *et al.* (1997). The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**(5331), 1453–1474.
49. Perna, N. T., Plunkett, G., 3rd, Burland, V., Mau, B., Glasner, J. D., Rose, D. J. *et al.* (2001). Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature*, **409**(6819), 529–533.
50. Lan, R. & Reeves, P. R. (2000). Intraspecies variation in bacterial genomes: the need for a species genome concept. *Trends Microbiol.* **8**, 396–401.
51. Boucher, Y., Nesbo, C. L. & Doolittle, W. F. (2001). Microbial genomes: dealing with diversity. *Curr. Opin. Microbiol.* **4**, 285–289.
52. Cole, S. T., Eigimeier, K., Parkhill, J., James, K. D., Thomson, N. R., Wheeler, P. R. *et al.* (2001). Massive gene decay in the leprosy bacillus. *Nature*, **409**, 1007–1011.
53. Andersson, S. G., Zomorodipour, A., Andersson, J. O., Sicheritz-Ponten, T., Alsmark, U. C., Podowski, R. M. *et al.* (1998). The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature*, **396**(6707), 133–140.
54. Ogata, H., Audic, S., Renesto-Audiffren, P., Fournier, P. E., Barbe, V., Samson, D. *et al.* (2001). Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. *Science*, **293**(5537), 2093–2098.
55. Andersson, J. O. & Andersson, S. G. (2001). Pseudogenes, junk DNA and the dynamics of rickettsia genomes. *Mol. Biol. Evol.* **18**, 829–839.
56. Parkhill, J., Wren, B. W., Thomson, N. R., Titball, R. W., Holden, M. T., Prentice, M. B. *et al.* (2001). Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature*, **413**(6855), 523–527.
57. Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H. *et al.* (1996). Life with 6000 genes. *Science*, **274**, 546, 563–567.
58. Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T. *et al.* (1998). SGD: saccharomyces genome database. *Nucl. Acids Res.* **26**(1), 73–79.
59. Harrison, P. M., Kumar, A., Lan, N., Echols, N., Snyder, M. & Gerstein, M. (2002). A small reservoir of disabled ORFs in the sequenced yeast genome and its implications for the dynamics of proteome evolution. *J. Mol. Biol.* in press.
60. Liu, H., Styles, C. A. & Fink, G. R. (1996). *S. cerevisiae* S288C has a mutation in FLO8, a gene required for filamentous growth. *Genetics*, **144**, 967–978.
61. Serio, T. R. & Lindquist, S. L. (2000). Protein-only inheritance in yeast: something to get [PSI +]-ched about. *Trends Cell Biol.* **10**, 98–105.
62. Eaglestone, S. S., Cox, B. S. & Tuite, M. F. (1999). Translation termination efficiency can be regulated in *S. cerevisiae* by environmental stress through a prion-mediated mechanism. *EMBO J.* **18**, 1974–1981.
63. Tuite, M. F. (2000). Yeast prions and their prion-forming domain. *Cell*, **100**, 289–292.
64. True, H. L. & Lindquist, S. L. (2000). A yeast prion provides a mechanism for genetic variation and phenotypic diversity. *Nature*, **407**(6803), 477–483.
65. Chervitz, S. A., Aravind, L., Sherlock, G., Ball, C. A., Koonin, E. V., Dwight, S. S. *et al.* (1998). Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science*, **282**, 2022–2028.
66. *C. elegans* Sequencing Consortium, T. (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, **282**, 2012–2018.

67. Rubin, G. M., Yandell, M. D., Wortman, J. R., Gabor Miklos, G. L., Nelson, C. R., Hariharan, I. K. *et al.* (2000). Comparative genomics of the eukaryotes. *Science*, **287**, 2204–2215.
68. Gopal, S., Schroeder, M., Pieper, U., Sczyrba, A., Aytekin-Kurban, G., Bekiranov, S. *et al.* (2001). Homology-based annotation yields 1042 new candidate genes in the *Drosophila melanogaster* genome. *Nature Genet.* **27**(3), 337–340.
69. Harrison, P. M., Echols, N. & Gerstein, M. (2001). Digging for dead genes: an analysis of the characteristics and distribution of the pseudogene population in the *C. elegans* genome. *Nucl. Acids Res.* **29**, 818–830.
70. Robertson, H. M. (2000). 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. *Genome Res.* **10**(2), 192–203.
71. Bargmann, C. I. (1998). Neurobiology of the *Caenorhabditis elegans* genome. *Science*, **282**, 2028–2033.
72. Remm, M. & Sonnhammer, E. (2000). Classification of transmembrane protein families in the *Caenorhabditis elegans* genome and identification of human orthologs. *Genome Res.* **10**, 1679–1689.
73. Glusman, G., Yanai, I., Rubin, I. & Lancet, D. (2001). The complete human olfactory subgenome. *Genome Res.* **11**, 685–702.
74. Zozulya, S., Echeverri, F. & Nguyen, T. (2001). The human olfactory receptor repertoire. *Genome Biol.* **2**, research0018.0011–research0018.0012.
75. Robin, G. C. Q., Russell, R. J., Cutler, D. J. & Oakeshott, J. G. (2000). The evolution of an alpha-esterase pseudogene inactivated in the *Drosophila melanogaster* lineage. *Mol. Biol. Evol.* **17**, 563–575.
76. Currie, P. D. & Sullivan, D. T. (1994). Structure, expression and duplication of genes which encode phosphoglyceromutase of *Drosophila melanogaster*. *Genetics*, **138**, 353–363.
77. Sullivan, D. T., Starmer, W. H., Curtiss, S. W., Menotti-Raymond, M. & Yum, J. (1994). Unusual molecular evolution of an Adh pseudogene in *Drosophila*. *Mol. Biol. Evol.* **11**, 443–458.
78. Petrov, D. A., Lozovskaya, E. R. & Hartl, D. L. (1996). High intrinsic rate of DNA loss in *Drosophila*. *Nature*, **384**, 346–349.
79. Petrov, D. A. & Hartl, D. L. (2000). Pseudogene evolution and natural selection for a compact genome. *J. Heredit.* **91**, 221–227.
80. Petrov, D. A. (2001). Evolution of genome size: new approaches to an old problem. *Trends Genet.* **17**, 23–28.
81. Ranz, J. M., Casals, F. & Ruiz, A. (2001). How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus Drosophila. *Genome Res.* **11**, 230–239.
82. Robertson, H. M. (1998). Two large families of chemoreceptor genes in the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae* reveal extensive gene duplication, diversification, movement and intron loss. *Genome Res.* **8**, 449–463.
83. Robertson, H. M. (2001). Updating the str and srj (stl) families of chemoreceptors in *Caenorhabditis nematodes* reveals frequent gene movement within and between chromosomes. *Chem. Senses*, **26**(2), 151–159.
84. Zhang, Z., Harrison, P.M. & Gerstein, M. (2002). Digging deep for ancient relics: a survey of protein motifs in the intergenic DNA of four eukaryotic genomes. submitted
85. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G. *et al.* (2001). The sequence of the human genome. *Science*, **291**(5507), 1304–1351.
86. Liang, F., Holt, I., Pertea, G., Karamychea, S., Salzberg, S. & Quackenbush, J. (2000). GEne index analysis of the human genome estimates approximately 120,000 genes. *Nature Genet.* **24**, 239–240.
87. Dunham, I., Shimizu, N., Roe, B. A., Chissoe, S., Hunt, A. R., Collins, J. E. *et al.* (1999). The DNA sequence of human chromosome 22. *Nature*, **402**(6761), 489–495.
88. Hattori, M., Fujiyama, A., Taylor, T. D., Watanabe, H., Yada, T., Park, H. S. *et al.* (2000). The DNA sequence of human chromosome 21. The chromosome 21 mapping and sequencing consortium. *Nature*, **405**(6784), 311–319.
89. Crollius, H. R., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C. *et al.* (2000). Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nature Genet.* **25**, 235–238.
90. Ewing, B. & Green, P. (2000). Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genet.* **232**, 232–233.
91. Wright, F. A., Lemon, W. J., Zhao, W. D., Sears, R., Zhuo, D., Wang, J. P. *et al.* (2001). A draft annotation and overview of the human genome. *Genome Biol.* **2**(7).
92. Mironov, A. A., Fickett, J. W. & Gelfand, M. S. (1999). Frequent alternative splicing of human genes. *Genome Res.* **9**, 1288–1293.
93. Brett, D., Hanke, J., Lehmann, G., Haase, S., Delbruck, S., Krueger, S. *et al.* (2000). EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Letters*, **474**, 83–86.
94. Modrek, B., Resch, A., Grasso, C. & Lee, C. (2001). Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucl. Acids Res.* **29**(13), 2850–2859.
95. Yeh, R.-F., Lim, L. P. & Burge, C. (2001). Computational inference of homologous gene structures in the human genome. *Genome Res.* **11**, 803–816.
96. Harrison, P. M., Hegyi, H., Balasubramaniam, S., Luscombe, N. M., Bertone, P., Echols, N. *et al.* (2002). Molecular fossils in the human genome: Identification and analysis of the pseudogenes on chromosomes 21 and 22. *Genome Res.* in press.
97. Goncalves, I., Duret, L. & Mouchiroud, D. (2000). Nature and structure of human genes that generate retropseudogenes. *Genome Res.* **10**(5), 672–678.
98. Dudov, K. P. & Perry, R. P. (1984). The gene family encoding the mouse ribosomal protein L32 contains a uniquely expressed intron-containing gene and an unmutated processed gene. *Cell*, **37**(2), 457–468.
99. Weiner, A. M. (2000). Do all SINEs lead to LINEs? *Nature Genet.* **24**, 332–333.
100. Koch, A. L. (1972). Enzyme evolution I. The importance of untranslatable intermediates. *Genetics*, **72**, 297–316.
101. Trabesinger-Ruef, N., Jermann, T., Zankel, T., Durrant, B., Frank, G. & Benner, S. A. (1996). Pseudogenes in ribonuclease evolution: a source of new biomacromolecular function? *FEBS Letters*, **382**(3), 319–322.

102. Sharon, D., Glusman, G., Pilpel, Y., Khen, M., Gruetzner, F., Haaf, T. & Lancet, D. (1999). Primate evolution of an olfactory receptor cluster: diversification by gene conversion and recent emergence of pseudogenes. *Genomics*, **61**(1), 24–36.
103. Ota, T. & Nei, M. (1995). Evolution of immunoglobulin VH pseudogenes in chickens. *Mol. Biol. Evol.* **12**, 94–102.
104. Aravind, L., Watanabe, H., Lipman, D. J. & Koonin, E. V. (2000). Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc. Natl Acad. Sci. USA*, **97**, 11319–11324.
105. Lykke-Andersen, J. (2001). mRNA quality control: marking the message for life or death. *Curr. Biol.* **11**(3), R88–R91.

Edited by F.E. Cohen

(Received 14 September 2001; received in revised form 1 February 2002; accepted 2 February 2002)