doi:10.1006/jmbi.2001.5343 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2002) **316**, 409–419

JMB

COMMUNICATION

A Small Reservoir of Disabled ORFs in the Yeast Genome and its Implications for the Dynamics of Proteome Evolution

Paul Harrison[1], Anuj Kumar[2], Ning Lan[1], Nathaniel Echols[1] Michael Snyder[1,2] and Mark Gerstein[1]*

[1]*Department of Molecular Biophysics & Biochemistry, and*

[2]*Department of Molecular Cellular & Developmental Biology, Yale University, 266 Whitney Ave., P.O. Box 208114, New Haven CT 06520-8114, USA*

**Corresponding author*

We surveyed the sequenced *Saccharomyces cerevisiae* genome (strain S288C) comprehensively for open reading frames (ORFs) that could encode full-length proteins but contain obvious mid-sequence disablements (frameshifts or premature stop codons). These pseudogenic features are termed disabled ORFs (dORFs). Using homology to annotated yeast ORFs and non-yeast proteins plus a simple region extension procedure, we have found 183 dORFs. Combined with the 38 existing annotations for potential dORFs, we have a total pool of up to 221 dORFs, corresponding to less than ~3 % of the proteome. Additionally, we found 20 pairs of annotated ORFs for yeast that could be merged into a single ORF (termed a mORF) by read-through of the intervening stop codon, and may comprise a complete ORF in other yeast strains. Focussing on a core pool of 98 dORFs with a verifying protein homology, we find that most dORFs are substantially decayed, with ~90 % having two or more disablements, and ~60 % having four or more. dORFs are much more yeast-proteome specific than live yeast genes (having about half the chance that they are related to a non-yeast protein). They show a dramatically increased density at the telomeres of chromosomes, relative to genes. A microarray study shows that some dORFs are expressed even though they carry multiple disablements, and thus may be more resistant to nonsense-mediated decay. Many of the dORFs may be involved in responding to environmental stresses, as the largest functional groups include growth inhibition, flocculation, and the SRP/TIP1 family. Our results have important implications for proteome evolution. The characteristics of the dORF population suggest the sorts of genes that are likely to fall in and out of usage (and vary in copy number) in a strain-specific way and highlight the role of subtelomeric regions in engendering this diversity. Our results also have important implications for the effects of the [PSI+] prion. The dORFs disabled by only a single stop and the mORFs (together totalling 35) provide an estimate for the extent of the sequence population that can be resurrected readily through the demonstrated ability of the [PSI+] prion to cause nonsense-codon read-through. Also, the dORFs and mORFs that we find have properties (e.g. growth inhibition, flocculation, vanadate resistance, stress response) that are potentially related to the ability of [PSI+] to engender substantial phenotypic variation in yeast strains under different environmental conditions. (See genecensus.org/pseudogene for further information.)

Keywords: translation termination; bioinformatics; genome annotation; pseudogene; yeast strains

Abbreviations used: ORF, open reading frame; dORF, disabled ORF; mORF, merged ORF; NMD, nonsense-mediated decay; SCRT, stop-codon readthrough.

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

0022-2836/02/030409–11 $35.00/0

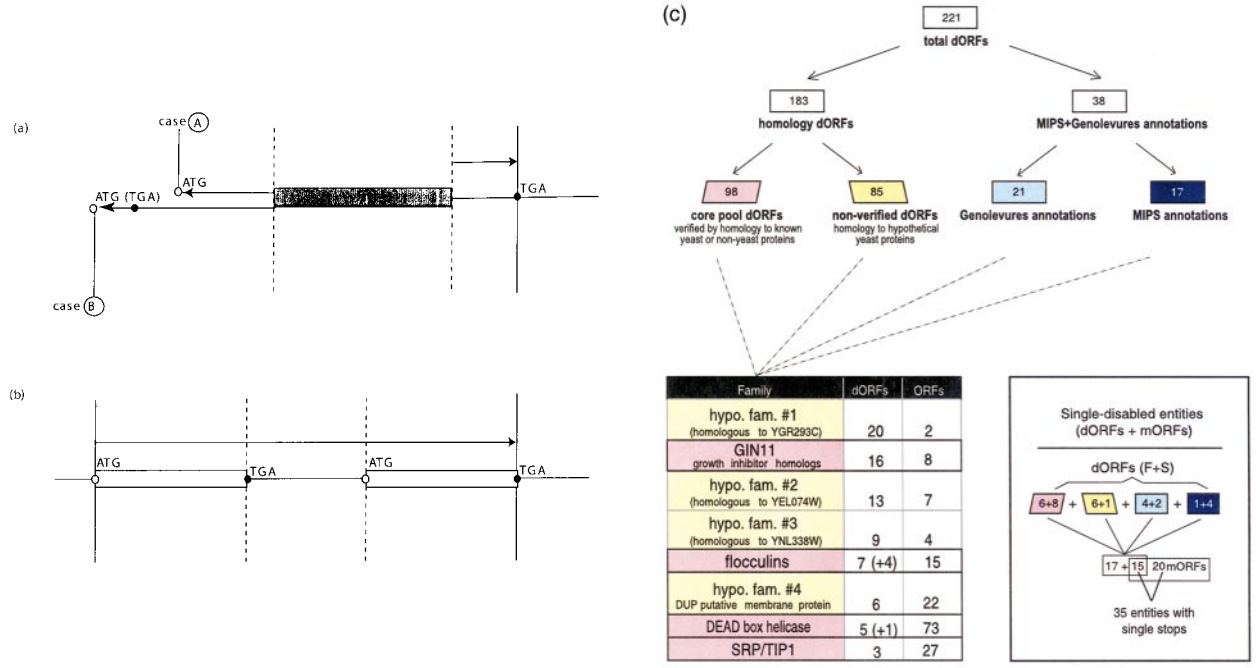

Figure 1 (*legend opposite*)

A disabled open reading frame (dORF) is defined as an ORF that is disabled by premature stop codons or frameshifts. Primarily, such dORFs are likely to be pseudogenes. Pseudogenes are ''dead'' copies of genes whose disablements imply that they do not form a full-length, functional protein chain. Two forms of pseudogenes generally occur: ''processed'' pseudogenes, where an mRNA transcript is reverse transcribed and re-integrated into the genome;[1] and ''non-processed'' pseudogenes, which arise from duplication of a gene in the genomic DNA and subsequent disablement.[2] Pseudogene populations have been described for human chromosomes 21 and 22, for the worm and for the prokaryotes *Mycobacterium leprae*, *Yersinia pestis* and *Rickettsia prowazekii*.[3–9] In the prokaryotes and in yeast, because of the shorter generation time such pseudogenes are likely to be ''strain-specific'', with proteins falling in and out of use because of environmental pressures peculiar to a particular strain. In yeast, there are no processed pseudogenes,[10] but there are a few documented pseudogenes that have presumably arisen from duplication (see MIPS and SGD databases[11,12]).

Apart from pseudogenes, dORFs with a single disablement may be examples of sequencing errors. Finally, dORFs with a single frameshift may arise as examples of +1 or −1 programmed ribosomal frameshifting. There is at present one verified example of either of these in the yeast genome.[13,14]

Determination of the extent and characteristics of the pool of dORFs in the sequenced yeast genome is important for furthering our understanding of yeast proteome evolution. Furthermore, it may shed light on effects of the [PSI+] prion on stop-codon read-through and the engendering of phenotypic diversity in yeast.[15]

Finding dORFs in the sequenced yeast genome

Since the full extent of the dORF complement in yeast is not known at present, here we have defined the yeast dORF pool using a simple homology-based procedure. As described in detail in Figure 1(a), the yeast genome was scanned for significant protein homologies that contain at least one disablement and that do not rely on alignment to a previously annotated ORF in the genomic DNA. That is, if the dORF entails an annotated ORF, the disabled extension to the ORF arises from a significant span of homology. The most appropriate dORF was then formed around each suitable disabled protein homology fragment (Figure 1(a)).

With our homology-based procedure, we find 183 dORFs. We also collated existing annotations of a further 38 dORFs and pseudogenic fragments from Genolevures hemi-ascomycete sequencing[16] and from MIPS[12] (17 from MIPS, 21 from Genolevures; Figure 1 and Table 1). This gives a grand total of up to 221 dORFs from all sources (Figure 1(c)). Of the 183 homology dORFs that we find, 98 (54 %) of them have verifying homology to either a known yeast protein or a non-yeast protein (Figure 2(b)). Known yeast proteins are those that have classes 1 through 3 in the MIPS ORF classification.[12] We focus on this core pool of 98 dORFs here as a verified set that was derived uniformly by a single procedure, setting aside those

dORFs that are homologous only to yeast hypothetical proteins and those based only on existing annotations. Core-pool dORFs with three or less disablements are given in Table 1, along with existing dORF annotations from the MIPS/Genolevures databases that could be discerned to have three or less disablements.

Additionally, we searched for pairs of existing annotated ORFs that are adjacent along the chromosome, and could be merged by stop codon read-through for the 5′ ORF of the pair, forming a single complete ORF (Figure 1(b)). We found 20 pairs of such merged ORFs, or mORFs (Table 2). One could consider this an additional method for finding dORFs with a single stop codon, but only those that arise from existing annotations, and that would form a whole ORF in a different yeast strain.

Properties of yeast dORFs

We examined the core pool of dORFs as follows: (1) their distribution of disablements; (2) their homology trends; (3) their prevalent families; and (4) their chromosomal distribution.

Disablements

Most dORFs are substantially decayed. The distribution of the number of disablements is shown

Figure 1. dORF and mORF detection. (a) dORFs from disabled protein homology. Initially, the complete sequenced genome of yeast[31] was searched in six-frame translation against the SWISSPROT protein sequence database[32] and yeast proteome sequence data from SGD (http://genome-www.stanford.edu/Saccharomyces, downloaded May 2000) and MIPS (http://mips.gsf.de, downloaded May 2000), using the alignment program TFASTX/Y.[33] Low complexity was masked using SEG.[34] All protein matches that overlapped genomic features such as transposable elements and tRNA genes were deleted. All significant protein matches (*e*-value $\leqslant 0.01$) were reduced for overlap by selecting homology segments in decreasing order of significance and flagging any others that overlap them for deletion. Matched stretches of genomic DNA that contained any disablements (either frameshifts or stop codons) were then further examined by comparing to the matching protein, a larger segment of the genomic DNA that had been extended at either end by the size of the matching protein sequence (in the equivalent number of nucleotides). This was performed with the FASTX/Y program. These enlarged homology fragments (denoted by the grey box) were then extended into the most appropriate ORFs, by searching for the nearest downstream stop codon (filled dot, TGA given as an example), and the farthest upstream start codon (open dot, labelled ATG at position A), or failing that, the nearest upstream start codon, after the nearest upstream stop (shown at position B). All such generated ORFs were then inspected manually, and reduced for overlap with each other where a larger predicted dORF comprises a similar shorter one. After this initial search for dORFs, we performed a second more comprehensive search for homology using PSI-BLAST.[35] We extracted all possible ORFs of size $\geqslant 30$ codons from the yeast genome (i.e. all stretches of genomic DNA beginning with start codon and ending with a stop codon) and searched them (in translation) against SWISSPROT[32] plus the combined annotated proteomes of *Caenorhabditis elegans*,[36] *Arabidopsis thaliana*,[37] *Drosophilia melanogaster*,[38] *S. cerevisiae* itself and 18 prokaryotes. All significant protein matches (using default threshold values) were again selected and processed as above for the original searches to find additional dORFs, again using FASTX/Y in the re-alignment stage. Those found only with PSIBLAST are labelled in Table 1. To gather existing annotation on potential dORFs, we examined the MIPS database[12] for any annotated pseudogenes, or ORFs reported to have stop codons or frameshifts. Also, from the Genolevures hemi-ascomycete sequencing project,[16] there are 17 examples of ORF extensions that may be potential dORFs (five singly disabled) that were not found by our disabled homology-searching procedure. Generally, these ORF extensions could be sequencing errors or (strain-specific) pseudogenes. All dORFs were checked against yeast chromosome sequence updates at http://genome-www.stanford.edu/Saccharomyces, resulting in the deletion of one dORF from the list. All yeast ORF classifications are taken from the MIPS database as of May 2000; known ORFs are those with classes 1 through 3. Sequencing to estimate stop codon errors: putative disablements were verified experimentally within all six non-repetitive and previously unidentified dORFs possessing a single premature stop codon. For purposes of this analysis, genomic DNA was extracted from a derivative of *S. cerevisiae* strain S288C. A region of this DNA encompassing each predicted premature stop codon was amplified using the polymerase chain reaction (PCR); PCR-amplified products were subsequently sequenced on both strands by standard methods (i.e. cycle-sequencing using big-dye terminators). By this approach, the presence of each premature termination codon was confirmed unambiguously. (b) Mergeable pairs of ORFs (mORFs). All adjacent pairs of annotated ORFs in the yeast genome (denoted by white boxes) were assessed for whether the 5′ partner of the pair could merge into the 3′ partner if the stop codon of the former is changed to a sense codon. If the two ORFs can form a larger ORF, ignoring the intervening stop codon, then the complete disabled reading frame is termed a mORF. (c) Classification of dORFs and mORFs. In the top panel, the tree shows the breakdown of the grand total of 221 dORFs into the 183 homology dORFs that we detected by our procedure, and the 38 additional annotations for dORFs and pseudogenic fragments culled from the MIPS and Genolevures databases.[12,16] The 183 homology dORFs separate into 98 dORFs with a verifying homology to a non-yeast protein or to a known yeast protein, and 85 dORFs that are homologous only to hypothetical ORFs. The inset panel at the bottom right describes the breakdown of the entities with single disablements (both dORFs and mORFs). Here, the dORFs for each of the four main groupings are shown with boxes of the same colour as in the top panel. The totals are split (frameshifts plus stops). The dORFs with single stops combined with the 20 mORFs give a total of 35 entities with single stop-codons. The inset panel at the bottom left shows the families in the dORF pool that have three or more members, with their corresponding numbers of ORFs. dORF families were derived using a modification of the algorithm described earlier.[39–41] dORFs and ORFs were deemed related if they have an alignment score of 1×10^{-4} or less for BLASTP.[35]

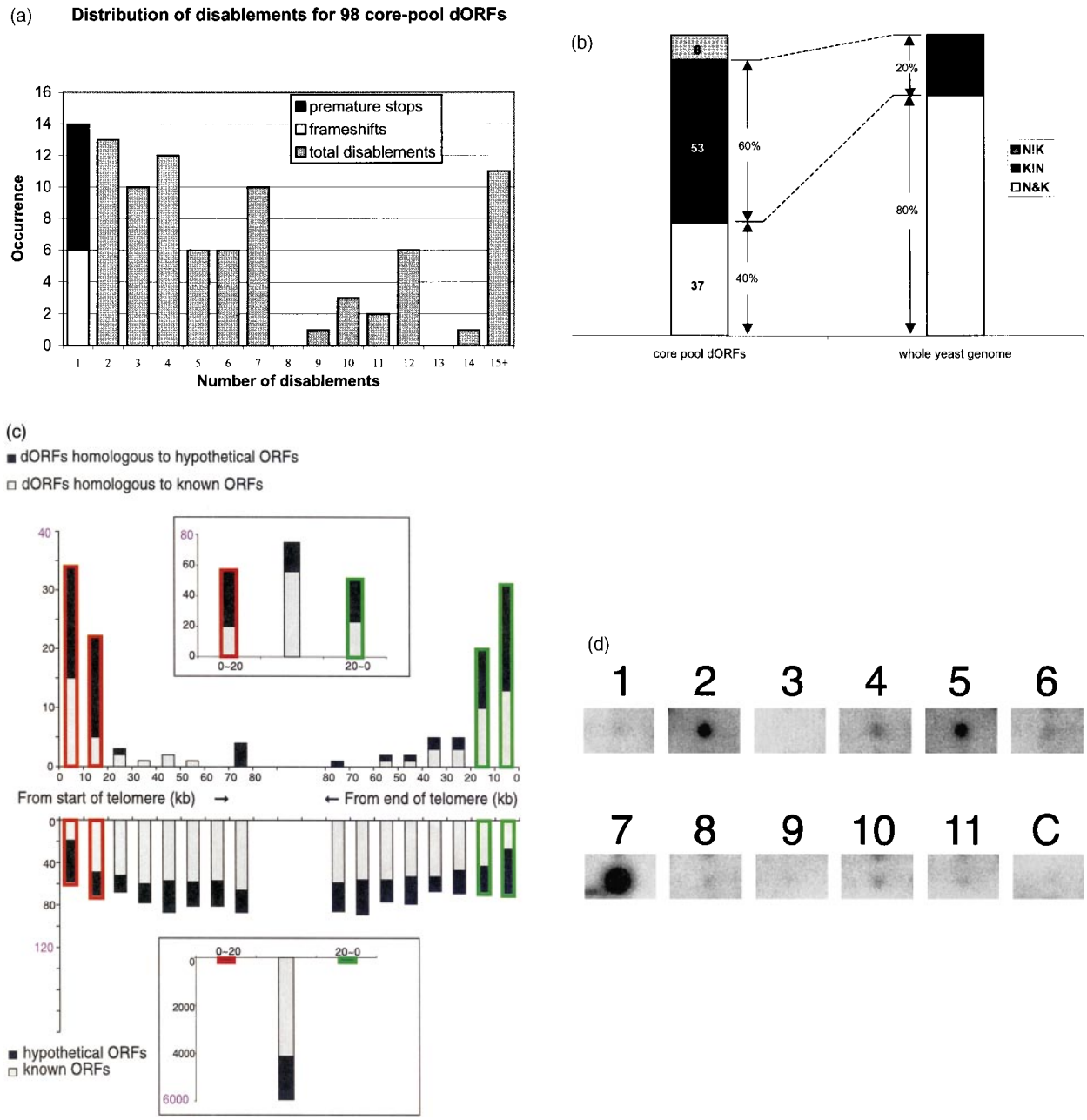

Figure 2 (*legend opposite*)

for the core pool of dORFs (in Figure 2(a)); 61% (60/98) have four or more disablements. In this set, 14 dORFs have one disablement, and eight of these a single premature stop codon (Table 1). An additional seven dORFs that are homologous only to hypothetical yeast proteins have a single disablement (one with a premature stop).

The existence of dORFs with single stop codons could be of relevance to the effects of the [PSI+] prion. Therefore, we checked the dORFs that we found by re-sequencing them (described in the legend to Figure 1(a)). We were able to amplify PCR products for six dORFs that were in non-repetitive regions, and verified the premature stop codons for each of them.

Homology trends

For some insight into strain-specific variation, we looked in more detail at the homology relationships of the 98 core-pool dORFs. Over half (54%) of these dORFs are specific to the *Saccharomyces cerevisiae* species, having no homology to non-yeast proteins (Figure 2(b)).

Four-fifths of the known yeast proteins (MIPS ORF classes 1 to 3[12]) are homologous to a non-yeast protein. In comparison, only about two-fifths

(41 %) of the dORFs that are homologous to a known yeast protein are homologous also to a non-yeast protein (Figure 2(b)). These homology trends change only slightly (±2 %) upon inclusion of the dORFs and pseudogenic fragments from the MIPS and Genolevures databases.

Furthermore, from the grand total of 221 dORFs, there are only a small number of dORFs (11) that correspond to ''live'' ORFs with no living relatives. One example is a very decayed reading frame of the KSH killer toxin corresponding to the single live KSH copy in the proteome (this protein also has no orthologs).

Prevalent families

Families of dORFs with three or more members are listed (Figure 1(c)). The family related to the growth inhibitor GIN11[17] stands out as the largest (16 members). The large population of growth-inhibitor dORFs may indicate that these vary in copy number for different yeast strains. The next largest family is the flocculins. These proteins have a variety of roles related to cell-cell adhesion, and are involved in mating, invasive growth and pseudohyphal formation in response to environmental stresses.[18] Pseudogenes for these have been discussed.[19] Most important of these is FLO8, which has a single stop-codon mutation in the laboratory strain S288C that prevents flocculation and filamentous growth (Table 1).[20] There are also five DEAD-box helicase dORFs (which is an abundant ORF family in yeast, Figure 1(c)) and three for the SRP/TIP1 family, which are involved in environmental stress response.

Highly increased density of dORFs at telomeres

We observe a highly increased density of dORFs at the telomeres of the chromosomes (Figure 2(c)). Out of our core pool of 98 verified dORFs, 43 (44 %) are subtelomeric, i.e. in the first and last 20 kb of the chromosomes. These include all of the dORFs for the two largest families, the flocculins and growth inhibitors noted in the previous section. If the 38 additional MIPS and Genolevures annotations are included, the proportion of dORFs in these telomeric intervals drops slightly (to 36 %). An even larger number of dORFs occur in the subtelomeric regions that are homologous only to hypothetical proteins (64 in the first and last 20 kb of the chromosomes out of the total of 85 non-verified dORFs that we find). Also, a quarter (5/20) of the mORFs are in the first and last 20 kb of the chromosomes. In comparison, the proportion of total gene annotations in these 20 kb telomeric

Figure 2. Analysis of the dORF reservoir. (a) The distribution of the number of disablements. This is shown for the core pool of 98 verified dORFs. The total for singly disabled dORFs is divided into those with a single frameshift (dark bar) and those with a single premature stop codon (white bar). The total disablements for ''15+'' includes all those counts greater than 15. Additionally, as can be deduced from Table 1, eight out of the 21 Genolevures-derived dORFs have at least four disablements. Disablements for the MIPS-annotated dORFs are not determined readily, as some of them are ORF truncations and pairs of homology fragments that would not be detected by our procedure. Those for which we could define the number of disablements are listed in Table 1. (b) Homology classification of dORFs. The distribution of the dORFs into those that have a non-yeast proteome homolog but no known yeast protein homolog (denoted N!K in Table 1), those that have a known yeast protein homolog but no non-yeast homolog (denoted K!N), and those that have both (denoted N&K). Inclusion of the homology trends for the extra MIPS and Genolevures annotations changes the representation of these categories only slightly (±2 % at most). (c) Highly increased density of dORFs at the telomeres. Distribution of dORFs (top panel) and ORFs (bottom panel) at the telomeres *versus* the remainder of the yeast chromosomes. The total number of dORFs and ORFs are shown in 10 kb intervals from both ends of all 16 yeast chromosomes (totalled together). In the bottom panel, known ORFs are shown with grey bars and hypothethical ORFs are shown with black ones. In the top panel, dORFs are divided into those homologous only to hypothetical yeast proteins (black bars) and the remainder (grey bars). The inset graphs show the total number of dORFs (upper panel) and ORFs (lower panel) within 20 kb from both telomeres and in the remaining span of the chromosomes. (d) Detected expression of dORFs. To investigate expression of dORF sequences, a sampling of 11 predicted dORFs were subjected to dot blot analysis using strand-specific oligonucleotides in an array-based format. For this analysis, poly(A) RNA was extracted from a vegetatively growing diploid S288C derivative; extracted RNA was treated with DNase I and subsequently biotinylated using the BrightStar™ Psoralen-Biotin kit (Ambion, Austin, TX). Biotinylated RNA was used to probe an array of 50-60-mer oligonucleotides spotted onto a nylon membrane-coated glass slide (Schleicher and Schuell, Keene, NH). Oligonucleotide sequences were derived from each putative dORF coding region and were selected to avoid repeated segments. Arrayed oligonucleotides were hybridized against 200 ng of biotinylated poly(A) RNA supplemented with denatured salmon sperm DNA at a final concentration of 100 μg/ml. Hybridizations were carried out in buffer containing formamide at 45(C. Bound RNA was detected using the BrightStar™ BioDetect™ kit (Ambion, Austin, TX). Spot size and intensity were quantified using software distributed in the NIH Image package version 1.62 (rsb.info.nih.gov/nih-image). Four dORF transcripts detected at levels appreciably distinct from background are shown here (lanes 2, 4, 5 and 7). These are homologs of the yeast ORFs YGR293c, YNL338w, YIL058w and YKL221w respectively. Lane C (negative control) indicates a lack of observable binding associated with hybridization against a non-coding region of the yeast genome. This dot blot analysis cannot be used to distinguish between transcripts greater than 75 % identical. As lane 2 and lane 4 are each representative of larger dORF families, this analysis indicates that at least one dORF from each of these previously unappreciated families is expressed under conditions of vegetative growth.

Table 1. dORFs with three or fewer disablements

Identifier[a]	Chromosome	Start	End	Sense	NK class[b]	Closest matching sequence (italic +[h]) or annotated genes involved (bold)[c]	Disablements[d]	Comment
A. *Core pool homology dORFs*								
D1-1	I	176,649	177,146	–	K!N	**YAR020C (PAU7)**	S	Involved in stress response; has stress-induced proteins SRP1/TIP1 family signature
D1-2	II	812,351	812,713	–	K!N	*YJR162C [h]*	S	Subtelomeric dORF belonging to large family related to Gin11 (a growth inhibitor)
D1-3	II	7605	8033	–	K!N	*YGL261C [h]*	F	Subtelomeric dORF in large family related to Gin11 (a growth inhibitor); has stress-induced proteins of SRP1/TIP1 family
D1-4	III	228,036	229,777	+	N&K	**YCR065W**	F	Transcription factor
D1-5	IV	1,527,751	1,527,939	+	K!N	**YDR545W**	F	Y′-helicase protein 1, from subtelomeric family of ORFs
D1-6	IX	439,074	439,345	–	K!N	*YJR162C [h]*	S	Homologous to Gin11 growth inhibitor
D1-7	V	176,580	176,795	+	K!N	*YDR366C [h]*	S	Has a protein-splicing motif
D1-8	VIII	215,187	217,899	–	K!N	**YHR056C**	F	Transcription regulator
D1-9	XV	2108	2651	–	K!N	*YCR106W [h]*	S	Transcription factor
D1-10	I	227,812	229,222	+	N&K	**YAR073W, YAR075W**	F	Similar to IMP dehydrogenase
D1-11	III	9324	11,147	+	N&K	**YCL069W**	S	Similar to drug resistance protein SGE1
D1-12	X	726,816	727,973	+	N&K	YJR155W	S	Similar to aryl-alcohol reductase
D1-13	X	31,866	32,150	+	N!K	*ADEC_ECOLI [h]*[e]	S	Adenine deaminase
D1-14	XIV	472,023	472,990	+	N&K	**YNL083W**	F	Extension to ORF YNL083W, similar to mitochondrial transport proteins
D2-1	II	6225	6600	+	K!N	*YJR162C [h]*	2	Similar to Gin11 protein
D2-2	III	830	1336	+	K!N	*YJR162C [h]*	2	Similar to Gin11 protein
D2-3	III	79,125	82,255	+	N&K	**YKL101W**	2	Ser/Thr-protein kinase involved in cell-cycle progression
D2-4	VI	990	2432	–	K!N	*YHR219W [h]*	2	Homologous to Y′-encoded proteins (DNA recombination)
D2-5	VII	531,242	531,531	+	K!N	*YOR196C [h]*	2[f]	Lipoic acid synthase
D2-6	X	117,956	119,581	–	K!N	**YJL160C**	2	Homologous to proteins involved in stress response; homologous to Pir1p/Hsp150p/Pir3p family
D2-7	XIII	5967	6346	+	K!N	*YJR162C [h]*	2[f]	Similar to Gin11 protein
D2-8	XV	1,083,930	1,084,380	–	K!N	*YFL063W [h]*	2	Member of the subtelomeric family involving Gin11
D2-9	XVI	942,413	942,642	–	K!N	*YNR077C [h]*	2	Member of the subtelomeric family involving Gin11
D2-10	XVI	6776	7224	+	K!N	*YFL063W [h]*	2	Member of the subtelomeric family involving Gin11
D2-11	III	100,944	101,291	–	N!K	*YEA3_SCHPO [h]*[e]	2	Hypothetical *Schizosaccharomyces pombe* protein
D2-12	VII	855,475	855,809	+	N!K	*YVFB_VACCC [h]*	2[f]	Hypothetical *Vaccinia* virus protein
D2-13	X	392,497	392,813	–	N!K	*YVFC_VACCC [h]*	2	Hypothetical *Vaccinia* virus protein
D3-1	III	293,055	293,261	+	K!N	*YCL066W [h]*	3[f]	Copy of HML mating type regulatory protein
D3-2	III	302,460	302,663	-	K!N	*YNR067C [h]*	3	Similar to β-glucan-elicitor receptor
D3-3	IX	428,922	429,214	+	N&K	*YER102W [h]*	3	Ribosomal protein S8e
D3-4	XII	1,064,292	1,065,175	-	K!N	*YFL063W [h]*	3	Part of subtelomeric family similar to Gin11
D3-5	XV	463,737	464,001	+	N&K	*YLR231C [h]*	3	Similar to kynureninase (involved in co-factor biosynthesis)
D3-6	III	108,713	110,292	+	N&K	**YCL004W**	3	Phosphatidylglycerophosphate synthase
D3-7	IV	768,472	769,204	-	N&K	**YML078W**	3[f]	Mitochondrial peptidyl prolyl isomerase
D3-8	X	237,531	237,838	-	N!K	*YVX3_CAEEL [h]*[e]	3	Hypothetical oxidoreductase
D3-9	VII	912,759	913,334	-	N&K	**YGR209C**	3	Decayed C-terminal extension to YGR209C (thioredoxin II)
D3-10	VII	936,017	936,446	-	N&K	**YGR220C**	3	Small extension (11 residues) to essential 50 S ribosomal protein L3P

Table 1 (*continued*).

Identifier	Disablements[d]	Comment
B. *MIPS annotations for dORFs*		
YFL051C	F	Similar to Flo (flocculin) genes, e.g. FLO10
YDR007W	S	TRP1, phosphoribosylanthranilate isomerase (tryptophan metabolism)
YOR031W	S	Metallothionein-like protein
YDR134C	S	Flocculin pseudogene
YER109C	S	FLO8, flocculin pseudogene; the gene is needed for diploid filamentous growth
YOL153C	2[f]	CPS1 (YJL172W) homolog (a carbosypeptidase)
C. *Genolevures annotations for potential dORFs*		
YJL213W	S	Similar to *Methanobacterium* aryldialkylphosphatase-related protein
YLR054C	S	Similar to *Saccharomyces bayanum* ORF
YBR041W	F	Fatty acyl-CoA synthetase
YGL059W	F	Similar to an α-keto-acid dehydrogenase kinase
YHR176W	F	Flavin-containing monooxygenase
YJL160C	F	Similar to Pir1p/Hsp150p/Pir3p family
YKR058W	2[f]	Initiator of glycogen synthesis
YMR207C	3[f]	Similar to acetyl CoA carboxylase
YNR062C	3[f]	Similar to *Hemophilus influenzae* lactate permease

If not from this work. Either MIPS or ''Geno'' (Genolevures hemi-ascomycete sequencing project).[16]

[a] This identifier is just for the purposes of this Table and is simply *D* plus the number of disablements plus a unique number.

[b] The NK class is given for each dORF found in the present survey and indicates whether the dORF is homologous to both non-yeast (N) and distinct known yeast (K) proteins (N&K), to non-yeast proteins but not known yeast (N!K), or to known yeast proteins, but not to non-yeast (K!N).

[c] The closest matching yeast protein for the dORFs (in italics); where the dORF encompasses a known ORF, its identifier is given in bold. For dORFs with no yeast homolog, a SWISSPROT identifier is given.

[d] For one disablement, it is specified whether there is a frameshift (F) or a premature stop (S). For more than one disablement, the number is indicated; in these cases this is the number of disablements for the homology segment around which the dORF is built.

[e] Found using PSI-BLAST,[35] as explained in Figure 1(a).

[f] These are dORFs that comprise only premature stop codons (no frameshifts).

Table 2. mORFs

ORF name	ORF name	Comment
YBR226C	YBR227C	qORF→clpx chaperone[a]
YDR504C	YDR505C	Hypo. protein→suppressor of ts mutations on DNA polymerase α
YDR082W	YDR083W	Involved in telomere length regulation→involved in rRNA processing
YDR157W	YDR158W	qORF→aspartate-semialdehyde dehydrogenase
YIL165C	YIL164C	Both homologous to parts of nitrilase[a]
YIR043C	YIR044C	Saccharopine dehydrogenase→COS family
YIL087C	YIL086C	Similar to hypo. *S. pombe* protein→hypo. protein
YIL168W	YIL167W	Both similar to serine dehydratase[a]
YER039C	YER039C-A	Similar to vanadate resistance protein Gog5→hypo. protein
YHR057C	YHR058C	Peptidyl-prolyl *cis-trans* isomerase→transcriptional regulation mediator
YKL031W	YL030W	Hypo. protein→qORF
YKL021C	YKL020C	MAK11, M1-virus replication protein→suppressor of Ty-induced promoter mutations
YKR032W	YKR034W	Hypo. protein→transcriptional repressor
YLR463C	YLR465C	Hypo. subtelomeric protein→qORF
YLR365W	YLR366W	Similar to Udf2p→hypo.protein
YMR056C	YMR057C	ADP/ATP carrier protein→hypo. mitochondrial protein
YNR068C	YNR069C	Both similar to Bul1p ubiquitination protein[a]
YOR024W	YOR025W	Hypo.protein→transcriptional silencing protein
YOR050C	YOR051C	Hypo.protein→weakly similar to myosins
YOL162W	YOL163W	Hypo.protein→phthalate transporter (both together are homologous to YLR004C)

All of the pairs are merged dORFs arising from the stop-codon read-through procedure in Figure 1(b). Hypo., hypothetical; qORF, questionable ORF as defined by MIPS;[20] →, precedes.
[a] Could be classed as dORFs.

intervals is very small (~4 %) (Figure 2(c)). These data indicate clearly the existence of a dynamically evolving subtelomeric subproteome in yeast.

Expression of dORFs

We tested a small random sample of 11 dORFs for expression (Figure 2(d)). Four of these showed appreciable expression, even though one has two disablements, and the other three have five or more disablements. Two of these four dORFs are subtelomeric (within 20 kb from chromosome ends), and homologous to putative hypothetical ORFs, representing dORF families of nine or more members. The other two are single dORFs with moderate sequence similarity for two annotated ORFs, both with five or more disablements; it is intriguing that we can still detect expression of these dORFs, an observation suggesting that these sequences, at minimum, possess functional promoters, and are still detectable despite nonsense-mediated decay (NMD).[21]

Implications for proteome evolution

A dynamically evolving subtelomeric subproteome and its role in strain-specific variation

The total pool of dORFs and pseudogenic fragments corresponds to only a very small percentage of the total annotated proteome (~3 %). However, the distribution of these dORFs, both in terms of homology and chromosomal position, details an important perspective on yeast proteome evolution.

In the present study, we have found that dORFs are half as likely to be related to a non-yeast protein (~40 % of dORFs) as to the average known yeast protein (80 % of annotated ORFs). This comparison implies that there has been no major change in the recent evolutionary dynamics of the yeast proteome. That is, it appears that disablement attacks evolutionarily young ORFs preferentially as opposed to ancient ORFs that are conserved between species. Also, there is a dramatically increased density of dORFs near the telomeres; as noted above, the two largest families of dORFs (flocculins and growth inhibitors) are subtelomeric and are related to subtelomeric ORFs. Additionally, a third interesting subtelomeric family that is classed as hypothetical but has a large number of dORFs (six compared to 21 live ORFs), is the DUP family of putative membrane proteins, which has an InterPro motif,[22] and whose expression may be pheromone-responsive.[23] It is interesting to note that subtelomeric regions can be meiotic recombination ''coldspots''.[24]

We have shown that some dORFs can still be expressed despite their disabled state, and may be more refractive to NMD in some way. This implies that such dORFs are still live to some extent, and represent a store of coding information.

Implications for the effects of the [PSI+] prion

[PSI+] is an inheritable phenomenon in yeast that is caused by the propagation of an alternatively folded, amyloid-like form of the Sup35p protein.[25,26] Sup35p is part of the surveillance complex in yeast that controls mRNA NMD and translation termination.[27] The occurrence of the [PSI+] prion in a yeast strain thus can lead to decreased translation termination efficiency as a result of stop-codon readthrough (SCRT), and increase the

likelihood that a protein will be formed from a dORF with a premature stop codon. SCRT for the *ade* gene has been used since the mid-1960 s as the standard protocol to detect the presence of [PSI+].[25,29] Different yeast strains show widely varied phenotypes for growth and viability in different environments depending on whether [PSI+] is present.[15,27] Thus, arguably, different levels of increased SCRT in yeast strains may be involved in causing this prion-engendered variability. It is possible that ribosomal frameshifting may be under the influence of the surveillance complex and consequently of [PSI+].[29] Although the sequenced yeast strain S288C is not a potent carrier of [PSI+], we examine below the size and make-up of our yeast dORF pool, particularly those that involve one stop codon, for the implications of [PSI+]-engendered phenotypic diversity in yeast.

The highest levels of [*PSI*+]-related SCRT for yeast strains that we can find in the literature are ~30 %,[27,29] with base-line levels in [*psi*−] cells of up to 5 %.[27,29] This implies that, assuming SCRT events are independent, ORFs with two or more stop codons are unlikely to produce substantial levels of encoded protein, even with [*PSI*+].

Consequently, we can use our data to estimate the size of the pool of sequence entities in a yeast strain that could be affected by SCRT caused by [PSI+]. We find that there is only a rather small cohort of 35 protein sequences that could be acted on readily by [PSI+] in this way. This comprises the set of all dORFs with a single premature stop codon, plus the mORFs that we detected (see the inset in Figure 1(c) for an explanation of this data set). This set of 35 entities corresponds to less than 1 % of the whole yeast proteome. Its small size suggests that minor extensions to existing annotated ORFs that are not detectable by homology may play a role in engendering phenotypic diversity in yeast.[15,27] On average, a yeast ORF would be extended by 17(±24) amino acid residues by SCRT; this may be long enough to add an additional secondary structure to a domain or a transmembrane helix.

The dORFs with a single stop codon (in Table 1), and the prevalent dORF families (Figure 1(c)) show characteristics that may be relevant to phenotypes arising from SCRT. As the presence of [PSI+] produces widely different growth phenotypes for different yeast strains, the number and state of decay of dORFs of the growth inhibitors (related to Gin11p) may have a bearing on [PSI+] strain-specific growth-rates.[15] The dORFs related to SRP stress-response proteins may have a role in cold-shock response. Of the single stop-codon dORFs that we observe, an extra viable copy of the fermentation enzyme aryl-alcohol reductase or of the drug-resistance pump SGE1 (Table 1) may prove beneficial for growth on different media. Finally, variation in flocculence (clumping from cell-cell adhesion) was observed in the recent study by True & Lindquist[15] on phenotypic diversity engendered by [PSI+]. Here, flocculins (which cause such cell-cell adhesion;[19]) comprise a large dORF family (Figure 1(c)), including three singly disabled dORFs. Variability in the number of distinct flocculins may help maintain a degree of strain-specific variation in cell adhesion properties. Flocculins are involved also in environmental stress response.[18]

We have detected mRNA transcripts corresponding to four dORFs possessing varying degrees of coding disability (Figure 2(d)). From this observation, we can suggest that the dORFs are real sequence entities and that disablements in coding sequence do not necessarily prohibit corresponding detectable mRNA sequence expression. The detected expression may imply that some dORFs are more refractive to NMD in some way, or may be interesting candidates for more detailed and comprehensive study of SCRT and the potential effects of [PSI+].

There are some interesting examples of mORFs that may have relevance for [*PSI*+] phenotypic diversity effects (Table 2). Note, however, that a large proportion of the ORFs involved (16/40) are hypothetical and that these MORFs may be complete ORFs in other yeast. For example:

YBR226C-YBR227C, a mitochondrial chaperone can be readthrough into from a hypothetical protein (predicted to be mitochondrial[30]); modification of the activity of this protein may affect mitochondrial protein homeostasis.

YHR057C-YHR058C, a peptidyl-prolyl isomerase can be N-terminally tagged onto a transcriptional regulation protein. These are clearly disparate functions; disruption of the latter ORF is lethal to yeast cells, so this fusion may decrease yeast-cell viability.

YER039C-YER039C-A, HVG1, which has strong similarity to vanadate-resistance protein (GOG5), can be readthrough into a short hypothetical protein (YER039C-A, 72 amino acid residues). This last pairing is particularly notable, since one yeast strain (with SCRT levels of ~26 %) showed a decreased growth-rate in the presence of vanadate when carrying [PSI+].[15] Also, HVG1 is the only paralog of GOG5 in the sequenced yeast strain S288C.

The mORFs we detected have linking nucleotide sequences of varying length (from one to 262 nucleotides, with a mean of 31). One could consider them as dORFs, but those that arise only from two existing ORF annotations; we assume that such mORFs could be complete ORFs in another yeast strain.

Website

The dORF annotation data and sequences are available at the website http://genecensus.org/pseudogene (or http://bioinfo.mbb.yale.edu/genome/pseudogene).

Acknowledgments

We thank Tricia Serio and Zhaolei Zhang for comments on the manuscript. A.K. is supported by a postdoctoral fellowship from the American Cancer Society. M.G. acknowledges support from the NIH protein structure initiative (P50 grant GM62413-01).

References

1. Vanin, E. F. (1985). Processed pseudogenes: characteristics and evolution. *Annu. Rev. Genet.* **19**, 253-272.
2. Mighell, A. J., Smith, N. R., Robinson, P. A. & Markham, A. F. (2000). Vertebrate pseudogenes. *FEBS Letters,* **468**, 109-114.
3. Andersson, S. G., Zomorodipour, A., Andersson, J. O., Sicheritz-Ponten, T., Alsmark, U. C. M., Podowski, R. M. *et al.* (1998). The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature,* **396**, 133-140.
4. 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**, 467-470.
5. 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**, 489-495.
6. 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**, 311-319.
7. 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.
8. Harrison, P. M., Hegyi, H., Balasubramaniam, S., Luscombe, N., 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.* **12**, 273-281.
9. 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.
10. Esnault, C., Maestre, J. & Heidmann, T. (2000). Human LINE retrotransposons generate processed pseudogenes. *Nature Genets.* **24**, 363-367.
11. 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**, 73-79.
12. 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**, 37-40.
13. Hammell, A. B., Taylor, R. C., Peltz, S. W. & Dinman, J. (1997). Identification of putative programmed −1 ribosomal frameshift signals in large DNA databases. *Genome Res.* **9**, 417-427.
14. Morris, D. K. & Lundblad, V. (1997). Programmed ribosomal frameshifting in a gene required for yeast telomere replication. *Curr. Biol.* **7**, 969-976.
15. True, H. L. & Lindquist, S. L. (2000). A yeast prion provides a mechanism for genetic variation and phenotypic diversity. *Nature,* **407**, 477-483.
16. Blandin, G., Durrens, P., Tekaia, F., Aigle, M., Bolotin-Fukuhara, M., Bon, E. *et al.* (2000). Genomic exploration of the hemiascomycetous yeasts: 4. The genome of *Saccharomyces cerevisiae* revisited. *FEBS Letters,* **487**, 31-36.
17. Kawahata, M., Amari, S., Nishizawa, Y. & Akada, R. (1999). A positive selection for plasmid loss in *S. cerevisiae* using galactose-inducible growth-inhibitory sequences. *Yeast,* **15**, 1-10.
18. Gancelo, J. M. (2001). Control of pseudohyphae formation in *Saccharomyces cerevisiae*. *FEMS Microbiol. Rev.* **25**, 107-123.
19. Teunissen, A. W. R. & Steensma, H. Y. (1995). The dominant flocculation genes of *Saccharomyces cerevisiae* constitute a new subtelomeric gene family. *Yeast,* **11**, 1001-1013.
20. Liu, H., Styles, C. A. & Fink, G. R. (1996). *Saccharomyces cerevisiae* S288C has a mutation in FLO8, a gene required for filamentous growth. *Genetics,* **144**, 967-978.
21. Lykke-Andersen, J. (2001). mRNA quality control: marking the message for life or decay. *Curr. Biol.* **11**, R88-R91.
22. 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**, 1145-1150.
23. Heiman, M. G. & Walter, P. (2000). Prm1p, a pheromone-regulated multispanning membrane protein, facilitates plasma membrane fusion during yeast mating. *J. Cell. Biol.* **151**, 719-730.
24. Gerton, J. L., DeRisi, J., Shroff, R., Lichten, M., Brown, P. O. & Petes, T. D. (2000). Global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA,* **97**, 11383-11390.
25. Serio, T. R. & Lindquist, S. L. (2000). Protein-only inheritance in yeast: something to get [PSI+]-ched about. *Trends Cell Biol.* **10**, 98-105.
26. Tuite, M. F. (2000). Yeast prions and their prion-forming domain. *Cell,* **100**, 289-292.
27. 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.
28. Cox, B. (1965). [PSI], a cytoplasmic suppressor of super-suppression in yeast. *Heredity,* **20**, 505-521.
29. Bidou, L., Stahl, G., Hatin, I., Namy, O., Rousset, J. P. & Farabaugh, P. J. (2000). Nonsense-mediated decay mutants do not affect programmed −1 frameshifting. *RNA,* **6**, 952-961.
30. Drawid, A. & Gerstein, M. (2001). A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *J. Mol. Biol.* **301**, 1059-1075.
31. Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H. *et al.* (1996). Life with 6000 genes. *Science,* **274**, 563-567.
32. Bairoch, A. & Apweiler, R. (2000). The SWISSPROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res.* **28**, 45-48.
33. Pearson, W. R., Wood, T., Zhang, Z. & Miller, W. (1997). Comparison of DNA sequences with protein sequences. *Genomics,* **46**, 24-36.

34. Wootton, J. C. & Federhen, S. (1996). Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266**, 554-571.
35. 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**, 3389-3402.
36. *C. elegans* Sequencing Consortium T. (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science,* **282**, 2012-2018.
37. Arabidopsis Genome Initiative T. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature,* **408**, 796-815.
38. Celniker, M. D., Holt, S. E., Evans, R. A., Gocayne, C. A., Amanatides, J. D., Scherer, P. G. *et al*. (2000). The genome sequence of *Drosophila melanogaster*. *Science,* **287**, 2185-2195.
39. Hobohm, U. & Sander, C. (1994). Enlarged representative set of protein structures. *Protein Sci.* **3**, 522-524.
40. Gerstein, M. (1997). A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J. Mol. Biol.* **274**, 562-576.
41. Hegyi, H. & Gerstein, M. (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288**, 147-164.

Edited by F. Cohen

(Received 26 June 2001; received in revised form 26 November 2001; accepted 26 November 2001)