A small reservoir of disabled ORFs in the *Saccharomyces cerevisiae* 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] & Mark B. Gerstein[1]

1: Dept. of Molecular Biophysics & Biochemistry,

2: Dept. of Molecular, Cellular & Developmental Biology,

Yale University,

266 Whitney Ave.,

P.O. Box 208114,

New Haven, CT 06520-8114,

U.S.A.

*Corresponding author
Phone: (203) 432-5065
Fax: (509) 691-6906
Email: harrison@csb.yale.edu

Submitted to JMB as a Communication, June 25th 2001

Revised version submitted, November 13th 2001

Summary

We comprehensively surveyed the sequenced *S. cerevisiae* genome (strain S288C) for open reading frames 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 get 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. 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 4 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. 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 readily 'resurrected' 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.

Keywords: translation termination, bioinformatics, genome annotation, pseudogene, yeast strains, prion

A 'disabled ORF' (dORF) is defined as an open reading frame 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 (Vanin, 1985); and 'non-processed' pseudogenes, which arise from duplication of a gene in the genomic DNA and subsequent disablement (Mighell et al., 2000). The 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* (Andersson, et al., 1998; Parkhill, et al., 2001) (Dunham et al., 1999; Hattori et al., 2000) (Harrison et al., 2001) (Harrison, et al., 2002, submitted) (Cole et al., 2001). 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 (Esnault et al., 2000), but there are a few documented pseudogenes that have presumably arisen from duplication (see MIPS and SGD databases; Cherry, et al., 1998; Mewes et al., 2000).

Apart from pseudogenes, dORFs with a single disablement may also 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 (Hammell et al., 1997) Morris & Lundblad, 1997).

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 the mechanism by effects of the [PSI+] prion on stop-codon read-through and the engendering of phenotypic diversity in yeast (True & Lindquist, 2000).

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 1a, 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 1a).

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 (Blandin et al., 2000) and from MIPS (Mewes et al., 2000) (17 from MIPS, 21 from Genolevures; Figure 1 legend and Table 1). This gives a grand total of up to 221 dORFs from all sources (Figure 1c). 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 2b). Known yeast proteins are those that have classes 1 through 3 in the MIPS ORF classification (Mewes, et al., 2000). We focus on this 'core pool' of 98 dORFs here as a verified set that was uniformly derived by a single procedure, setting aside those dORFs that are homologous only to yeast hypothetical proteins and those based only on existing annotations. Those from the core pool of dORFs with $\leq$ 3 disablements are listed in Table 1, along with existing dORF annotations from the MIPS / Genolevures databases that could be discerned to have $\leq$ 3 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 1b). We found twenty pairs of such merged ORFs, or 'mORFs' (Table 2).

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.

(1) Disablements. Most dORFs are substantially decayed. The distribution of the number of disablements is shown for the core pool of dORFs (in Figure 2a); 61% (60/98) have $\geq$4 disablements. In this set, there are 14 of these dORFs with one disablement, 8 of these with a single premature stop codon (Table 1). An additional 7 dORFs that are only homologous 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 using sequencing (described in Figure 1a legend). 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.

(2) 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 *S.cerevisiae* species, having no homology to non-yeast proteins (Figure 2b).

Four-fifths of the known yeast proteins (MIPS ORF classes 1 to 3; Mewes, et al., 2000) 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 *also* homologous to a non-yeast protein (Figure 2b). 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 (eleven) 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).

(3) Prevalent families. Families of dORFs with three or more members are listed (Figure 1c). The family related to the growth inhibitor GIN11 (YLL065W; Kawahata et al., 1999) 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 (Gancedo, 2001). Pseudogenes for these have been discussed previously (Teunissen & Steensma, 1995). 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) Liu et al., 1996). There are also five DEAD-box helicase dORFs (which is an abundant ORF family in yeast, Figure 1c) and three for the SRP/TIP1 family, which are involved in environmental stress response.

(4) Highly increased density of dORFs at telomeres. We observe a highly increased density of dORFs at the telomeres of the chromosomes (Figure 2c). 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%). There is an even larger number of dORFs occurring in the subtelomeric regions that are homologous only to hypothetical proteins (64 in the first and last 20 kilobases 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 intervals is very small (~4%) (Figure 2c). This data clearly indicates the existence of a dynamically evolving subtelomeric subproteome in yeast.

Expression of dORFs

We tested a small random sample of eleven dORFs for expression (Figure 2d). Four of these showed appreciable expression, even though one has two disablements, and the other three have ≥5 disablements. Two of these four dORFs are subtelomeric (within 20 kb from chromosome ends), and homologous to putative hypothetical ORFs, representing dORF families of size ≥9 members. The other two are single dORFs with moderate sequence similarity for two annotated ORFs, both with ≥5 disablements----it is intriguing that we can still detect expression of these dORFs, an observation suggesting that these sequences, at minimum, possess functional promoters.

Implications for proteome evolution

(1) 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 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 preferentially attacks evolutionarily young ORFs 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 (6 compared to 21 'live' ORFs), is the 'DUP' family of putative membrane proteins, which has an InterPro motif (Apweiler et al., 2000), and whose expression may be pheromone-responsive (Heiman & Walter, 2000). The pronounced concentration of subtelomeric dORFs is also consistent with subtelomeric regions as more recombinogenic regions (McEachern & Iyer, 2001), with increased recombination causing increased occurrence of disablements. The 'live' and 'dead' members of these subtelomeric families evidently form a rapidly evolving subproteome in yeast. Recombination has been demonstrated to be a generator or flocculin diversity (Kobayashi, *et al.*, 1998).

We have shown that some dORFs can still expressed despite their disabled state. This implies that such dORFs are still 'live' to some extent, represent a store of coding information, in the aftermath of a recombination event that has lead to disablement.

(2) 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 (Serio & Lindquist, 2000; Tuite, 2000). Sup35p is part of the surveillance complex in yeast that controls nonsense-mediated mRNA decay and translation termination (Eaglestone et al., 1999). The occurrence of the [PSI+] prion in a yeast strain thus can lead to decreased translation termination efficiency as a result of stop-codon read-through (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+] (Cox, 1965; Serio & Lindquist, 2000). Different yeast strains show widely varied phenotypes for growth and viability in different environments depending on whether or not [PSI+] is present (True & Lindquist, 2000; Eaglestone, et al., 1999). Thus, arguably, different levels of increased SCRT in yeast strains may be involved in causing this prion-engendered variability. It is also possible that ribosomal frameshifting may be under the influence of the surveillance complex and consequently of [PSI+] (Bidou et al., 2000). 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 [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% (Bidou et al., 2000; Eaglestone et al., 1999), with base-line levels in [*psi-*] cells of up to 5% (Bidou et al., 2000; Eaglestone et al., 1999). This implies that, assuming SCRT events are independent, ORFs with $\geq$2 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 readily acted on 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 Figure 1c inset 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 also play a role in engendering phenotypic diversity in yeast (True & Lindquist, 2000; Eaglestone, et al., 1999). 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 1c) 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 (True & Lindquist, 2000). 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 also prove beneficial for growth on different media. Finally, variation in flocculence (clumping from cell-cell adhesion) was observed in the recent study by True and Lindquist (True & Lindquist, 2000) on phenotypic diversity engendered by [PSI+]. Here, flocculins (which cause such cell-cell adhesion; see, *e.g.* (Teunissen & Steensma, 1995)) comprise a large dORF family (Figure 1c), including 3 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 also involved in environmental stress response (Gancedo, 2001).

We have detected mRNA transcripts corresponding to four dORFs possessing varying degrees of coding disability (Figure 2d). From this observation, we suggest that the dORFs are real sequence entities and that disablements in coding sequence do not necessarily prohibit corresponding sequence expression at the RNA level. Furthermore, this expression data indicate dORFs that 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; however a large proportion of the ORFs involved (16/40) are hypothetical). For example:-

YBR226c-YBR227c: a mitochondrial chaperone can be read-through into from a hypothetical protein (predicted to be mitochondrial; Drawid & Gerstein, 2000); disruption 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 read-through into a short hypothetical protein (YER039C-A, 72 amino acids). This last pairing is particularly notable since one yeast strain (with SCRT levels of ~26%) showed decreased growth rate in the presence of vanadate when carrying [PSI+] (True & Lindquist, 2000). 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 1 to 262 nucleotides, with a mean of 31). Two of the mORFs are probably better classed as dORFs, as the two merged sequences form a complete copy of a known protein (labelled in Table 2)

Website

The dORF annotation data and sequences are available at the website http://bioinfo.mbb.yale.edu/genome/pseudogene/yeast.

Acknowledgements

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).

Table 1: dORFs with 3 or fewer disablements

(a) ‘core pool’ homology dORFs

Identifier ¶¶¶	Chromosome	start	End	sense	NK class*	Closest matching sequence *(italic +[h])* or annotated genes involved (bold) **	Disablements ***	Comment
D1-1	I	176649	177146	-	K ! N	**YAR020C (PAU7)**	S	involved in stress response ; has stress-induced proteins SRP1/TIP1 family signature
D1-2	II	812351	812713	-	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	228036	229777	+	N&K	**YCR065W**	F	transcription factor
D1-5	IV	1527751	1527939	+	K ! N	**YDR545W**	F	Y’-helicase protein 1, from subtelomeric family of ORFs
D1-6	IX	439074	439345	-	K ! N	*YJR162C [h]*	S	homologous to Gin11 growth inhibitor
D1-7	V	176580	176795	+	K ! N	*YDR366C [h]*	S	has a protein-splicing motif
D1-8	VIII	215187	217899	-	K ! N	**YHR056C**	F	transcription regulator
D1-9	XV	2108	2651	-	K ! N	*YCR106W [h]*	S	transcription factor
D1-10	I	227812	229222	+	N&K	**YAR073W,**	F	similar to IMP-dehydrogenase

						YAR075W		
D1-11	III	9324	11147	+	N&K	**YCL069W**	S	similar to drug resistance protein SGE1
D1-12	X	726816	727973	+	N&K	**YJR155W**	S	similar to aryl-alcohol reductase
D1-13	X	31866	32150	+	N ! K	*ADEC_ECOLI [h]* ¶¶	S	adenine deaminase
D1-14	XIV	472023	472990	+	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	79125	82255	+	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	531242	531531	+	K ! N	*YOR196C [h]*	2¶	lipoic acid synthase
D2-6	X	117956	119581	-	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¶	similar to Gin11 protein
D2-8	XV	1083930	1084380	-	K ! N	*YFL063W [h]*	2	member of the subtelomeric family involving Gin11
D2-9	XVI	942413	942642	-	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	100944	101291	-	N ! K	*YEA3_SCHPO [h]* ¶¶	2	hypothetical *S. pombe* protein
D2-12	VII	855475	855809	+	N ! K	*YVFB_VACCC [h]*	2¶	hypothetical *Vaccinia* virus protein
D2-13	X	392497	392813	-	N ! K	*YVFC_VACCC [h]*	2	hypothetical *Vaccinia* virus protein

D3-1	III	293055	293261	+	K ! N	*YCL066W [h]*	3¶	copy of HML mating type regulatory protein
D3-2	III	302460	302663	-	K ! N	*YNR067C [h]*	3	similar to beta-glucan-elicitor receptor
D3-3	IX	428922	429214	+	N&K	*YER102W [h]*	3	ribosomal protein S8e
D3-4	XII	1064292	1065175	-	K ! N	*YFL063W [h]*	3	part of subtelomeric family similar to Gin11
D3-5	XV	463737	464001	+	N&K	*YLR231C [h]*	3	similar to kynureninase (involved in co-factor biosynthesis)
D3-6	III	108713	110292	+	N&K	**YCL004W**	3	phosphatidylglycerophosphate synthase
D3-7	IV	768472	769204	-	N&K	**YML078W**	3¶	mitochondrial peptidyl prolyl isomerase
D3-8	X	237531	237838	-	N ! K	*YVX3_CAEEL [h]*¶¶	3	hypothetical oxidoreductase
D3-9	VII	912759	913334	-	N&K	**YGR209C**	3	Decayed C-terminal extension to YGR209C (thioredoxin II)
D3-10	VII	936017	936446	-	N&K	**YGR220C**	3	Small extension (11 residues) to essential 50S ribosomal protein L3P

(b) MIPS annotations for dORFs

Identifier	Disablements***	Comment
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¶	CPS1 (YJL172W) homolog (a carbosypeptidase)

(c) Genolevures annotations for potential dORFs

Identifier	Disablements***	Comment
YJL213W	S	similar to *Methanobacterium* aryldialkylphosphatase-related protein
YLR054C	S	similar to *S. bayanum* ORF
YBR041W	F	Fatty acyl-CoA synthetase
YGL059W	F	Similar to an alpha-keto-acid dehydrogenase kinase
YHR176W	F	Flavin-containing monooxygenase
YJL160C	F	Similar to Pir1p/ Hsp150p/Pir3p family

YKR058W	2¶	Initiator of glycogen synthesis
YMR207C	3¶	Similar to acetyl CoA carboxylase
YNR062C	3¶	Similar to *H.influenzae* lactate permease

* 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).

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.

***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, if no specific start or stop points could be determined.

****If not from this work. Either MIPS or ‘Geno’ (Genolevures hemi-ascomycete sequencing project; see (Blandin et al., 2000)

¶ These are dORFs that comprise only premature stop codons (no frameshifts).

¶¶ Found using PSI-BLAST (Altschul et al., 1997), as explained in Figure 1a.

¶¶¶ This identifier is just for the purposes of this table and is simply *D* plus the number of disablements plus a unique number.

Table 2: mORFs*

ORF name	ORF name	Comment
YBR226C	YBR227C	qORF→ clpx chaperone
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 **
YIR043C	YIR044C	Saccharopine dehydrogenase → COS family
YIL087C	YIL086C	Similar to hypo. *S.pombe* protein → hypo. protein
YIL168W	YIL167W	Both similar to serine dehydratase **
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
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 1b. The following abbreviation/symbols are used: ‘hypo.’ = hypothetical, ‘qORF’ = questionable ORF as defined by MIPS (Mewes, et al., 2000), ‘→’ = precedes.

** Could also be classed as dORFs.

Figure legends

Figure 1: dORF and mORF detection.

(a) dORFs from disabled protein homology. Initially, the complete sequenced genome of yeast (Goffeau et al., 1996) was searched in six-frame translation against the SWISSPROT protein sequence database (Bairoch & Apweiler, 2000) 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 (Pearson et al., 1997). Low complexity was masked using SEG (Wootton & Federhen, 1996). All protein matches that overlapped genomic features such as transposable elements and tRNA genes were deleted. All significant protein matches (e-value $\leq$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 (black dot, TGA given as an example), and the farthest upstream start codon (unfilled 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 (Altschul et al., 1997). We extracted all possible ORFs of size $\geq$ 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 (Bairoch & Apweiler, 2000) plus the combined annotated proteomes of *C. elegans* (C. elegans Sequencing Consortium, 1998), *A. thaliana* (Arabidopsis Genome Initiative, 2000), *D. melanogaster* (Adams et al., 2000), *S. cerevisiae* itself and eighteen 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 (Mewes et al., 2000) for any annotated pseudogenes, or ORFs reported to have stop codons or frameshifts. Also, from the Genolevures hemi-ascomycete sequencing project (Blandin et al., 2000), there are 17 examples of ORF extensions that may be potential dORFs (5 singly-disabled) that were not found by our disabled homology-searching procedure. Generally, these ORF extensions could either be sequencing errors or be (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 experimentally verified 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 unambiguously confirmed.

(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 read through. If the two ORFs can form a larger ORF, ignoring the intervening stop codon, then the complete disabled reading frame is termed a mORF (for 'merged ORF').

(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 (Mewes, et al., 2000; Blandin, et al., 2000). 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 only homologous to hypothetical ORFs. The inset panel at 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 bottom left shows the families in the dORF pool that have 3 or more members, with their corresponding numbers of ORFs. dORF families were derived using a modification of the algorithm of Hobohm and Sander (1996). dORFs and ORFs were deemed related if they have an alignment score of $1x10^{-4}$ or less for BLASTP (Altschul et al., 1997).

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 as well. Additionally, as can be deduced from Table 1, eight out of the 21 *Genolevures*-derived dORFs have ≥4 disablements. Disablements for the MIPS-annotated dORFs are not readily determined, 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 only homologous 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^{TM} 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 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^{TM} BioDetect^{TM} 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.

References

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F. & Venter, J. C. (2000). The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185-2195.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**(17), 3389-402.

The genome sequence of Rickettsia prowazekii and the origin of mitochondria

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

Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M. D., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N. J., Oinn, T. M., Pagni, M., Servant, F., Sigrist, C. J. & Zdobnov, E. M. (2000). InterPro--an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* **16**(12), 1145-50.

Arabidopsis Genome Initiative, T. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796-815.

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

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**(7), 952-61.

Blandin, G., Durrens, P., Tekaia, F., Aigle, M., Bolotin-Fukuhara, M., Bon, E., Casaregola, S., de Montigny, J., Gaillardin, C., Lepingle, A., Llorente, B., Malpertuy, A., Neuveglise, C., Ozier-Kalogeropoulos, O., Perrin, A., Potier, S., Souciet, J., Talla, E., Toffano-Nioche, C., Wesolowski-Louvel, M., Marck, C. & Dujon, B. (2000). Genomic exploration of the hemiascomycetous yeasts: 4. The genome of Saccharomyces cerevisiae revisited. *FEBS Lett* **487**(1), 31-6.

Blattner, F. R., Plunkett, G., 3rd, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B. & Shao, Y. (1997). The complete genome sequence of Escherichia coli K-12. *Science* **277**(5331), 1453-74.

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

Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S. & Botstein, D. (1998). SGD: Saccharomyces Genome Database. *Nucleic Acids Res* **26**, 73-9.

Cole, S. T., Eigimeier, K., Parkhill, J., James, K. D., Thomson, N. R., Wheeler, P. R., Honore, N., Garnier, T., Churcher, C., Harris, D. & Barrell, B. G. (2001). Massive gene decay in the leprosy bacillus. *Nature* **409**, 1007-1011.

Cox B. (1965). [PSI], a cytoplasmic suppressor of super-suppression in yeast. *Heredity* **20**, 505-21

Dunham, I., Shimizu, N., Roe, B. A., Chissoe, S., Hunt, A. R., Collins, J. E., Bruskiewich, R., Beare, D. M., Clamp, M., Smink, L. J., Ainscough, R., Almeida, J. P., Babbage, A., Bagguley, C., Bailey, J., Barlow, K., Bates, K. N., Beasley, O., Bird, C. P., Blakey, S., Bridgeman, A. M., Buck, D., Burgess, J., Burrill, W. D., O'Brien, K. P. & et al. (1999). The DNA sequence of human chromosome 22. *Nature* **402**(6761), 489-95.

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.

Ehrenhofer-Murray, A. E., Seitz, M. U. & Sengstag, C. (1998). The Sge1 protein of Saccharomyces cerevisiae is a membrane-associated multidrug transporter. *Yeast* **14**(1), 49-65.

Esnault, C., Maestre, J. & Heidmann, T. (2000). Human LINE retrotransposons generate processed pseudogenes. *Nature Genets.* **24**, 363-367.

Gancelo, J.M. (2001). Control of pseudohyphae formation in Saccharomyces cerevisiae. *FEMS Microbiol. Revs.* **25**, 107-123.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H. & Oliver, S. G. (1996). Life with 6000 genes. *Science* **274**(5287), 546, 563-7.

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.

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.

Hattori, M., Fujiyama, A., Taylor, T. D., Watanabe, H., Yada, T., Park, H. S., Toyoda, A., Ishii, K., Totoki, Y., Choi, D. K., Soeda, E., Ohki, M., Takagi, T., Sakaki, Y., Taudien, S., Blechschmidt, K., Polley, A., Menzel, U., Delabar, J., Kumpf, K., Lehmann, R., Patterson, D., Reichwald, K., Rump, A., Schillhabel, M. & Schudy, A. (2000). The DNA sequence of human chromosome 21. The chromosome 21 mapping and sequencing consortium. *Nature* **405**(6784), 311-9.

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.

Hobohm, U. & Sander, C. (1994). Enlarged representative set of protein structures. *Protein Sci.* **3**, 522-4 .

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.

Kobayashi, O., Hayashi, N., Kuroki, R. & Sone, H. (1998). Region of Flo1 Proteins Responsible for Sugar Recognition. *J. Bacteriol.*, **180**, 6503-6510.

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.

McEachern, M. J. & Iyer, S. (2001). Short telomeres in yeast are highly recombinogenic. *Mol. Cell* **7**, 695-704.

Mewes, H. W., Frishman, D., Gruber, C., Geier, B., Haase, D., Kaps, A., Lemcke, K., Mannhaupt, G., Pfeiffer, F., Schuller, C., Stocker, S. & Weil, B. (2000). MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* **28**(1), 37-40.

Mighell, A. J., Smith, N. R., Robinson, P. A. & Markham, A. F. (2000). Vertebrate pseudogenes. *FEBS Letts* **468**, 109-114.

Morris, D. K. & Lundblad, V. (1997). Programmed ribosomal frameshifting in a gene required for yeast telomere replication. *Curr. Biol.* **7**, 969-976.

Parkhill J, Wren BW, Thomson NR, Titball RW, Holden MT, Prentice MB, Sebaihia M, James KD, Churcher C, Mungall KL, Baker S, Basham D, Bentley SD, Brooks K, Cerdeno-Tarraga AM, Chillingworth T, Cronin A, Davies RM, Davis P, Dougan G, Feltwell T, Hamlin N, Holroyd S, Jagels K, Karlyshev AV, Leather S, Moule S, Oyston PC, Quail M, Rutherford K, Simmonds M, Skelton J, Stevens K, Whitehead S, Barrell BG. (2001). Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature***, 413,** 467-470.

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

Perna, N. T., Plunkett, G., 3rd, Burland, V., Mau, B., Glasner, J. D., Rose, D. J., Mayhew, G. F., Evans, P. S., Gregor, J., Kirkpatrick, H. A., Posfai, G., Hackett, J., Klink, S., Boutin, A., Shao, Y., Miller, L., Grotbeck, E. J., Davis, N. W., Lim, A., Dimalanta, E. T., Potamousis, K. D., Apodaca, J., Anantharaman, T. S., Lin, J., Yen, G., Schwartz, D. C., Welch, R. A. & Blattner, F. R. (2001). Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. *Nature* **409**(6819), 529-33.

Serio, T. R. & Lindquist, S. L. (2000). Protein-only inheritance in yeast: something to get [PSI+]-ched about. *Trends Cell Biol* **10**(98-105).

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.

True, H. L. & Lindquist, S. L. (2000). A yeast prion provides a mechanism for genetic variation and phenotypic diversity. *Nature* **407**(6803), 477-83.

Tuite, M. F. (2000). Yeast prions and their prion-forming domain. *Cell* **100**, 289-292.

Vanin, E. F. (1985). Processed pseudogenes: characteristics and evolution. *Annu. Rev. Genets.* **19**, 253-272.

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