Published in final edited form as:
Science. 1997 August 29; 277(5330): 1259–1260.

Yeast as a Model Organism

David Botstein, **Steven A. Chervitz**, and **J. Michael Cherry**
Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA

David Botstein: botstein@genome.stanford.edu

The first complete DNA sequence of a eukaryotic genome, that of the yeast *Saccharomyces cerevisiae*, was released in electronic form more than a year ago (1). No doubt, each member of the international consortium of yeast biologists made the argument to his or her own funding agency in Europe, Japan, Britain, Canada, or the United States that this yeast would be a fine "model organism," useful for interpreting and understanding human DNA sequences. How right were they?

It was clear long before the systematic sequencing of genomes began that there are genes in yeast and mammals that encode very similar proteins (2). Some homologies—including proteins of molecular systems (for example, the ribosomes and cytoskeletons)—were no surprise. Some were quite unexpected, however. A particularly arresting early example was the discovery in yeast of two close homologs (*RAS1* and *RAS2*) of the mammalian *ras* proto-oncogene; yeast cells lacking both genes are inviable. In 1985 this system was the occasion for the first of many deliberate tests of functional conservation: The mammalian H-*ras* sequence was expressed in a yeast strain lacking both RAS genes, with the remarkable result that viability was restored, indicating a profound conservation not only of sequence, but also of detailed biological function (3).

With the entire yeast genome sequence in hand, we can estimate how many yeast genes have significant mammalian homologs. We compared (4) all yeast protein sequences to the mammalian sequences in GenBank [EST (expressed sequence tag) databases were not included]. The result (see the table) is encouraging: For nearly 31% of all the potential protein-encoding genes of yeast (open reading frames, or ORFs), we found a statistically robust homolog among the mammalian protein sequences (5). This is clearly an underestimate, as the databases surely do not yet contain the sequences of all mammalian proteins or even representatives of every protein family. Many of these similarities relate individual domains, and not whole proteins, no doubt reflecting the shuffling of functional domains characteristic of protein evolution.

Even though *S. cerevisiae* is among the best-studied experimental organisms, 60% of its genes still have no experimentally determined function. Of these, the majority nevertheless have some similarity or motif suggesting possible functions, leaving about 25% (by actual count) with no clue whatever. In compiling the data in the table, we observed that genes with homology to mammalian sequences are much less likely to have nothing experimental known of their function. Only 34% of the entire set of yeast genes with mammalian homologs have no function listed in the *Saccharomyces* Genome Database; compared to less than 25% of the genes having the strongest homology. We do not know the reason for this, although we do not rule out the optimistic idea that yeast biologists have succeeded in concentrating on the most important genes (those most likely to be conserved).

The likelihood that a newly discovered human gene will have a yeast homolog with at least some functional information about one of its domains is thus quite good. Genetic manipulation in yeast is easy and cheap, whereas such manipulation, even when possible in

mammalian systems, is neither easy nor cheap. There is in addition the opportunity to exploit functional compatibility by the method described above for the RAS genes. At least 71 human genes complement yeast mutations; this is certain to be an underestimate (6). Thus, information about human genes learned from studying their yeast homologs comes at an excellent price.

Probably the best examples of the value of yeast as a model system concern human disease genes that have been mapped by linkage, positionally cloned, and then sequenced. Usually nothing is known of these genes beyond the fact that their inheritance results in disease. The sequence of the gene generally provides the first clue to function by way of homology to the genes of other organisms, commonly *S. cerevisiae* (7). Among the best matches are the human genes that cause hereditary nonpolyposis colon cancer (*MSH2* and *MLH1* in yeast), neurofibromatosis type 1 (*IRA2* in yeast), ataxia telangiectasia (*TEL1* in yeast), and Werner's syndrome (*SGS1* in yeast). Two of these have particularly illustrative stories.

Inherited nonpolyposis colon cancers have a cellular phenotype: instability of short repeated sequences in the tumor cells. Stimulated by this result, and even before the human genes had been cloned, yeast researchers isolated mutations in yeast genes with the same phenotype (including mutations in *MSH2* and *MLHI*), predicting that the colon cancer genes were likely to be their homologs (8).

Werner's syndrome is a disease with several hallmarks of premature aging. Again there is a cellular phenotype, which includes a reduced life-span in culture. The sequence of the human gene was found to be highly similar to that of the yeast *SGS1* gene, which encodes a DNA helicase. On page 1313 of this issue, Sinclair *et al.* (9) report that *SGS1* mutant yeast cells have a markedly reduced life-span and share other cellular phenotyes with cells from individuals with Werner's syndrome.

So yeast has indeed turned out to be a useful "model" for eukaryotic biology. There is ample justification for intensifying efforts to determine the functional roles of the remaining 60% of yeast genes whose function is still not known. There are as well many individual reasons to focus even more attention on genes such as *MSH2* and *SGS1*. These yeast genes may represent the most efficient path to understanding the colon cancer and the aging caused by mutations in their human homologs.

References and Notes

1. Goffeau A, et al. Science 1996;274:546. *Saccharomyces* Genome Database (SGD) at http://genome-www.stanford.edu/Saccharomy-ces/; Yeast Genome from MIPS (Martinsried Institute for Protein Sequences) at http://speedy.mips.biochem.mpg.de/mips/yeast/; Yeast Protein Database (YPD) at http://www.proteome.com/YPDhome.html. [PubMed: 8849441]
2. Botstein D, Fink GR. Science 1988;240:1439. [PubMed: 3287619]
3. Kataoka T, et al. Cell 1985;40:19. [PubMed: 2981628]
4. BLASTP analysis were done between all yeast ORF translations and all unique protein sequences in the human, mouse, rat, cow, and sheep sequences in GenBank as of 22 July 1997. We used the BLOSUM62 substitution matrix and low-complexity filters seg and xnu. "Unknown function" means that the ORF had no entry in either the Gene_Product or Description fields within its SGD Locus page as of 30 July 1997. For all ORFs, 3783 (60.8%) have unknown function by this definition. BLASTP, version 2.0a, W. Gish, unpublished data; Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. J. Mol. Biol 1990;215:403. [PubMed: 2231712]
5. For details, see http://genome-www.stanford.edu/Saccharomyces/mammal/.
6. These data are in XREFdb from the National Center for Biotechnology Information at http://www.ncbi.nlm.nih.gov/Bassett/cerevisiae/index.html.
7. These data are in XREFdb at http://www.ncbi.nlm.nih.gov/Bassett/Yeast/PosiClonSceNew.html.

8. Strand M, Prolla TA, Liskay RM, Petes TD. Nature 1993;365:274. [PubMed: 8371783]
9. Sinclair DA, Mills K, Guarente L. Science 1997;277:1313. [PubMed: 9271578]
10. The *Saccharomyces* Genome Database is supported by an NIH research resources grant (HG 01315).

Table 1

Mammalian homologs (based on *P* value)

***P* value**	**Number of ORFs at *P* value or lower**	**Percent of total ORFs (*n* = 6223)**	**Percent of ORFs with unknown function**
1×10^{-10}	1914	30.8	34
1×10^{-20}	1553	25.0	30
1×10^{-40}	1083	16.8	26
1×10^{-60}	784	12.6	23
1×10^{-80}	576	9.3	22
1×10^{-100}	442	7.1	21
1×10^{-150}	221	3.6	23
1×10^{-200}	101	1.6	25