by FISH across the whole genome. A subset of the results were selected in which there was no evidence of the YAC having been chimeric, and the position was resolved to a cytogenetic interval (as opposed to fractional length only). A Généthon marker was given for each of these YACs, which allowed the position in centimorgans to be determined. Because this study contained no data for chromosomes 19, 21, and 22, an alternative strategy was used: Genes that are nonrecombinant with respect to Généthon markers [table 5 in C. Dib *et al.*, *Nature* **380** (suppl.), iii (1996)] and have well-known cytogenetic locations [Online Mendelian Inheritance in Man (OMIM) at http://www.ncbi.nlm.nih.gov/] were used to establish the cross-reference.
28. N. E. Morton, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 7474 (1991).
29. Detailed instructions and examples are provided on the Web site.
30. G. D. Schuler, J. A. Epstein, H. Ohkawa, J. A. Kans, *Methods Enzymol.* **266**, 141 (1996).
31. Five additional assays yielded ambiguous results as a result of the presence of an interfering mouse band or poor amplification.
32. This could be the result of either a trivial primer tube labeling error or a sequence clustering error in which two separate genes were erroneously assigned to the same UniGene entry.
33. Typing errors in a framework marker would allow a close EST to map with significant lod scores (logarithm of the odds ratio for linkage) to a correct chromosome location but would tend to localize the marker in a distant bin, in order to minimize "double-breaks" caused by the erroneous typings of the framework marker.
34. J. M. Craig and W. A. Bickmore, *Bioessays* **15**, 349 (1993).
35. V. Romano *et al.*, *Cytogenet. Cell Genet.* **48**, 148 (1988).
36. S. Bahram, M. Bresnahan, D. E. Geraghty, T. Spies, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 6259 (1994).
37. G. D. Billingsley *et al.*, *Am. J. Hum. Genet.* **52**, 343 (1993); S. S. Schneider *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 3147 (1995).
38. R. Sherrington *et al.*, *Nature* **375**, 754 (1995); D. Levitan and I. Greenwald, *ibid.* **377**, 351 (1995); S. A. Hahn *et al.*, *Science* **271**, 350 (1996).
39. S. Tugendreich *et al.*, *Hum. Mol. Genet.* **3**, 1509 (1994); D. E. Bassett Jr., M. S. Boguski, P. Hieter, *Nature* **379**, 589 (1996); P. Hieter, D. E. Bassett Jr., D. Valle, *Nature Genet.* **13**, 253 (1996).
40. The scoring systems were amino acid substitution matrices based on the PAM (point accepted mutation) model of evolutionary distance. Pam matrices may be generated for any number of PAMs by extrapolation of observed mutation frequencies [M. O. Dayhoff *et al.*, in *Atlas of Protein Sequence and Structure*, M. O. Dayhoff, Ed. (National Biomedical Research Foundation, Washington, DC, 1978), vol. 5, suppl. 3, pp. 345–352; S. F. Altschul, *J. Mol. Evol.* **36**, 290 (1993)]. PAM matrices were customized for scoring matches between sequences in each of the 15 species pairs; for the pool of remaining proteins ("Other organisms" in Table 4), the PAM120 matrix was used because it has been shown to be good for general-purpose searching [S. F. Altschul, *J. Mol. Biol.* **219**, 555 (1991)]. The BLASTX program [W. Gish and D. J. States, *Nature Genet.* **3**, 266 (1993)] takes a nucleotide sequence query (EST or gene), translates it into all six conceptual ORFs, and then compares these with protein sequences in the database; TBLASTN performs a similar function but instead searches a protein query sequence against six-frame translations of each entry in a nucleotide sequence database. Searches were performed with *E* = 1e-6 and *E2* = 1e-5 as the primary and secondary expectation parameters.
41. M. A. Pericak-Vance *et al.*, *Am. J. Hum. Genet.* **48**, 1034 (1991); W. J. Strittmatter *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 1977 (1993); R. Fishel *et al.*, *Cell* **75**, 1027 (1993); N. Papadopoulos *et al.*, *Science* **263**, 1625 (1994); R. Shiang *et al.*, *Cell* **78**, 335 (1994); L. M. Mulligan *et al.*, *Nature* **363**, 458 (1993); P. Edery *et al.*, *ibid.* **367**, 378 (1994).
42. The potential effect of a comprehensive gene map is well illustrated by the fact that 82% of genes that have been positionally cloned to date are represented by one or more ESTs in GenBank (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_genes/).
43. K. H. Fasman, A. J. Cuticchia, D. T. Kingsbury, *Nucleic Acids Res.* **22**, 3462 (1994).
44. M. Bonaldo *et al.*, *Genome Res.* **6**, 791 (1996).
45. Consensus sequences from at least three overlapping ESTs were generated from 294 UniGene clusters. Of the 230 (78%) redesigned primer pairs, 188 (82%) of these yielded successful PCR assays.
46. We thank M. O. Anderson, A. J. Collymore, D. F. Courtney, R. Devine, D. Gray, L. T. Horton Jr., V. Kouyoumjian, J. Tam, W. Ye, and I. S. Zemsteva from the Whitehead Institute for technical assistance. We thank W. Miller, E. Myers, D. J. Lipman, and A. Schaffer for essential contributions toward the development of UniGene. Supported by NIH awards HG00098 to E.S.L., HG00206 to R.M.M., HG00835 to J.M.S., and HG00151 to T.C.M., and by the Whitehead Institute for Biomedical Research and the Wellcome Trust. T.J.H. is a recipient of a Clinician Scientist Award from the Medical Research Council of Canada. D.C.P. is an assistant investigator of the Howard Hughes Medical Institute. The Stanford Human Genome Center and the Whitehead Institute–MIT Genome Center are thankful for the oligonucleotides purchased with funds donated by Sandoz Pharmaceuticals. Généthon is supported by the Association Francaise contre les Myopathies and the Groupement d'Etudes sur le Genome. The Sanger Centre, Généthon, and Oxford are grateful for support from the European Union EVRHEST programme. The Human Genome Organization and the Wellcome Trust sponsored a series of meetings from October 1994 to November 1995 without which this collaboration would not have been possible.

Life with 6000 Genes

A. Goffeau,* B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, S. G. Oliver

The genome of the yeast *Saccharomyces cerevisiae* has been completely sequenced through a worldwide collaboration. The sequence of 12,068 kilobases defines 5885 potential protein-encoding genes, approximately 140 genes specifying ribosomal RNA, 40 genes for small nuclear RNA molecules, and 275 transfer RNA genes. In addition, the complete sequence provides information about the higher order organization of yeast's 16 chromosomes and allows some insight into their evolutionary history. The genome shows a considerable amount of apparent genetic redundancy, and one of the major problems to be tackled during the next stage of the yeast genome project is to elucidate the biological functions of all of these genes.

The genome of the yeast *Saccharomyces cerevisiae* has been completely sequenced through an international effort involving some 600 scientists in Europe, North America, and Japan. It is the largest genome to be completely sequenced so far (a record that we hope will soon be bettered) and is the first complete genome sequence of a eukaryote. A number of public data libraries compiling the mapping information and nucleotide and protein sequence data from each of the 16 yeast chromosomes (*1–16*) have been established (Table 1).

The position of *S. cerevisiae* as a model eukaryote owes much to its intrinsic advantages as an experimental system. It is a unicellular organism that (unlike many more complex eukaryotes) can be grown on defined media, which gives the experimenter complete control over its chemical and physical environment. *S. cerevisiae* has a life cycle that is ideally suited to classical genetic analysis, and this has permitted construction of a detailed genetic map that defines the haploid set of 16 chromosomes. Moreover, very efficient techniques have been developed that permit any of the 6000 genes to be replaced with a mutant allele, or completely deleted from the genome, with absolute accuracy (*17–19*). The combination of a large number of chromosomes and a small genome size meant that it was possible to divide sequencing responsibilities conveniently among the different international groups involved in the project.

Old Questions and New Answers

The genome. At the beginning of the sequencing project, perhaps 1000 genes encoding either RNA or protein products had been defined by genetic analysis (*20*). The complete genome sequence defines some 5885 open reading frames (ORFs) that are likely to specify protein products in the yeast cell. This means that a protein-encoding gene is found for every 2 kb of the yeast genome, with almost 70% of the total sequence consisting of ORFs (*21*). The yeast genome is much more compact than those of its more complex relatives in the eukary-

otic world. By contrast, the genome of the nematode worm contains a potential protein-encoding gene every 6 kb (*22*), and in the human genome, some 30 kb or more of sequence must be examined in order to uncover such a gene. Analysis of the yeast genome reveals the existence of 6275 ORFs that theoretically could encode proteins longer than 99 amino acids. However, 390 ORFs are unlikely to be translated into proteins. Thus, only 5885 protein-encoding genes are believed to exist. In addition, the yeast genome contains some 140 ribosomal RNA genes in a large tandem array on chromosome XII and 40 genes encoding small nuclear RNAs (snRNAs) scattered throughout the 16 chromosomes; 275 transfer RNA (tRNA) genes (belonging to 43 families) are also widely distributed. Table 2, which provides details of the distribution of genes and other sequence elements among yeast's 16 chromosomes, shows that the genome has been completely sequenced, with the exception of a set of identical genes repeated in tandem.

The compact nature of the *S. cerevisiae* genome is remarkable even when compared with the genomes of other yeasts and fungi. Current data from the systematic sequence analysis of the genome of the fission yeast *Schizosaccharomyces pombe* indicate that the density of protein-encoding genes is approximately one per 2.3 kb (*23*). The difference between these two yeast genomes can be ascribed to the paucity of introns in *S. cerevisiae*. In the fission yeast, approximately 40% of genes contain introns (*21*), whereas only 4% of protein-encoding genes in *S. cerevisiae* are similarly interrupted (Table 2). The *Saccharomyces* genes that do contain introns [notably, those encoding ribosomal proteins (*24*)] usually have only one small intron close to the start of the coding sequence (often interrupting the initiator codon) (*25*). It has even been suggested (*26*) that many yeast genes represent cDNA copies that have been generated by the action of reverse transcriptases specified by retrotransposons (Ty elements).

The chromosomes. A complete genome sequence provides more information than the sum of all the genes (or ORFs) that it contains. In particular, it permits an investigation of the higher order organization of the *S. cerevisiae* genome. An example is the long-range variation in base composition. Many yeast chromosomes consist of alternating large domains of GC-rich and GC-poor DNA (*21*, *27*), generally correlating with the variation in gene density along these chromosomes. In the case of chromosome III, it has been demonstrated that the periodicity in base composition is paralleled by a variation in recombination frequency along the chromosome arms, with the GC-rich peaks coinciding with regions of high recombination in the middle of each arm of the chromosome and AT-rich troughs coinciding with the recombination-poor centromeric and telomeric sequences. Simchen and co-workers (*28*) have demonstrated that the relative incidence of double-strand breaks, which are thought to initiate genetic recombination in yeast (*29*), correlates directly with the GC-rich regions of this chromosome.

The four smallest chromosomes (I, III, VI, and IX) exhibit average recombination frequencies some 1.3 to to 1.8 times greater than the average for the genome as a whole. Kaback (*30*) has suggested that high levels of recombination have been selected for on these very small chromosomes to ensure at least one crossover per meiosis, and so permit them to segregate correctly. It is known that artificial chromosomes (or chromosome fragments) of approximately 150 kb in size are mitotically unstable (*31*, *32*), which raises the related questions of whether there is a minimal size for yeast chromosomes and how the smallest chromosomes have achieved their current size (see Table 2). The organization of chromosome I is very unusual: The 31 kb at each of its ends are very gene-poor, and Bussey *et al.* (*5*) have suggested that these terminal domains may act as "fillers" to increase the size, and hence the stability, of this smallest yeast chromosome.

Genetic redundancy is the rule at the ends of yeast chromosomes. For instance, the two terminal domains of chromosome III show considerable nucleotide sequence homology both to one another and to the terminal domains of other chromosomes (V and XI). The right terminal region of chromosome I is duplicated at the left end of chromosome I (*5*) and at the right end of chromosome VIII (*3*). The sugar fermentation genes *MAL*, *SUC*, and *MEL* all have a number of telomere-associated copies, not all of which are expressed (*33*–*35*). An interesting feature of the distribution of the

A. Goffeau and H. Tettelin, Université Catholique de Louvain, Unité de Biochimie Physiologique, Place Croix du Sud, 2/20, 1348 Louvain-la-Neuve, Belgium.
B. G. Barrell, Sanger Centre, Hinxton Hall, Hinxton, Cambridge CB10 1SA, UK.
H. Bussey, Department of Biology, McGill University, 1205 Docteur Penfield Avenue, Montreal, H3A 1B1, Canada.
R. W. Davis, Department of Biochemistry, Stanford University, Beckman Center, Room B400, Stanford, CA 94305–5307, USA.
B. Dujon, Unité de Génétique Moléculaire des Levures (URA1149 CNRS and UPR927 Université Pierre et Marie Curie), Institut Pasteur, 25 Rue du Docteur Roux, 75724 Paris-Cedex 15, France.
H. Feldmann, Universität München, Institut für Physiologische Chemie, Schillerstrasse, 44, 80336 München, Germany.
F. Galibert, Laboratoire de Biochimie Moleculaire, UPR 41–Faculté de Médecine, CNRS, 2 Avenue du Professeur Léon Bernard, 35043 Rennes Cedex, France.
J. D. Hoheisel, Molecular Genetic Genome Analysis, Deutsches Krebsforschungszentrum, Im Neuenheimer Feld, 506, 69120 Heidelberg, Germany.
C. Jacq, Génétique Moleculaire, Ecole Normale Supérieure, CNRS URA 1302, 46 Rue d'Ulm, 75230 Paris Cedex 05, France.
M. Johnston, Department of Genetics, Box 8232, Washington University Medical School, 4566 Scott Avenue, St. Louis, MO 63110, USA.
E. J. Louis, Yeast Genetics, Institute of Molecular Medicine, John Radcliffe Hospital, Oxford OX3 9DU, UK.
H. W. Mewes, Max-Planck-Institut für Biochemie, Martinsried Institute for Protein Sciences (MIPS), Am Klopferspitz 18A, 82152 Martinsried, Germany.
Y. Murakami, Tsukuba Life Science Center, Division of Human Genome Research, RIKEN, Koyaday Tsukuba Science City 3-1-1, 305 Ibaraki, Japan.
P. Philippsen, Institute for Applied Microbiology, Universität Basel, Klingelbergstrasse 70, 4056 Basel, Switzerland.
S. G. Oliver, Department of Biochemistry and Applied Molecular Biology, University of Manchester Institute of Science and Technology (UMIST), Post Office Box 88, Manchester M60 1QD, UK.
*To whom correspondence should be addressed.

Table 1. Finding yeast genome information on the Internet. A fuller description of these and other resources may be found in (*86*).

Internet addresses
FTP sites for the complete S. cerevisiae *genome sequence*
ftp.mips.embnet.org (directory/yeast)
ftp.ebi.ac.uk (directory/pub/databases/yeast)
genome-ftp.stanford.edu (directory/yeast/genome-seq)
Other S. cerevisiae *data libraries*
MIPS, Martinsried, Germany (http://www.mips.biochem.mpg.de/yeast/)
Sanger Center, Hinxton, UK (http://www.sanger.ac.uk/yeast/home.html)
Saccharomyces Genome Database (SGD), Stanford University, USA (http://genome-www.stanford.edu/Saccharomyces/)
SWISS-PROT, University of Geneva, Switzerland (http://expasy.hcuge.ch/sprot/sp-docu.html)
Yeast Protein Database (YPD), Proteome Inc., Beverly, MA, USA (http://www.proteome.com/YPDhome.html)
GeneQuiz, European Molecular Biology Laboratory, Heidelberg, Germany (http://www.embl-heidelberg.de/~genequiz/)
XREFdb, National Center for Biological Information, Baltimore, MD, USA (http://www.ncbi.nlm.nih.gov/XREFdb/)
(Editor's note: For readers who would like a user-friendly guide to the yeast databases, please see the special feature at the *Science* Web site http://www.sciencemag.org/science/feature/data/genomebase.htm)

MEL genes is that they are only found associated with one telomere of a chromosome and not both (*36*). This raises the intriguing possibility that yeast chromosomes might have an intrinsic polarity beyond our arbitrary labeling of left and right arms.

The left telomere of chromosome III has some special characteristics. Like all yeast telomeres, it contains a repeated sequence element called X (*37*, *38*). It also contains a pseudo–X element at an internal site about 4 kb from the true X, and Voytas and Boeke (*39*) have suggested that the two X sequences represent the long terminal repeats (LTRs) of a new class of yeast transposon called Ty5. Transposons are often found on the healed ends of broken chromosomes in *Drosophila* (*40*, *43*). The yeast genome sequence reveals that all 19 telomere-associated highly conserved repeats called Y′ elements contain an ORF whose predicted protein product is reminiscent of the RNA helicases (*42*, *43*). No Y′ helicase ORFs are found on chromosomes I, III, and XI (though there are small parts of Y′s on III and XI near the actual telomeres) and a few Y′s from tandem arrays at chromosomes XII R and IV R have not been sequenced. The functions of these ORFs are unknown; however, they may have formed parts of transposable elements in the past (*41*, *44*). Because the synthesis of new telomeric repeats by telomerase [see (*45*)] is essentially a reverse transcription process, it may be that current mechanisms of telomere biogenesis in many eukaryotes had their origins in the activities of retrotransposons or retroviruses.

Whatever the merits of such speculation, it is evident that many of the polymorphisms observed between homologous chromosomes in different strains of *S. cerevisiae* are due to transposition events or recombination between transposons or their LTRs or both (*46*). Fortunately for genome analysis, and perhaps for yeast itself, these spontaneous transposition events do not appear to occur randomly along the length of individual chromosomes. The yeast genome that was sequenced contains 52 complete Ty elements as well as 264 solo LTRs or other remnants that are the footprints of previous transposition events. The majority of the Ty*2* elements (11 out of 13) are found in sites that show evidence of previous transposition activity ("old" sites); only about half (16 out of 33) of the Ty*1* elements are found in "new" sites (the majority flanking tRNA genes). Thus, yeast transposons appear to insert preferentially into specific chromosomal regions that may be termed transposition hot spots (*46*–*53*).

The proteome. The term "proteome" has been coined to describe the complete set of proteins that a living cell is capable of synthesizing (*54*). The completion of the yeast genome sequence means that, for the first time, the complete proteome of a eukaryotic cell is accessible. Computer analysis of the yeast proteome allows classification of about 50% of the proteins on the basis of their amino acid sequence similarity with other proteins of known function, with the use of simple and conservative homology criteria. However, such assignments often provide only a general description of the biochemical function of the predicted protein products (such as "protein kinase" or "transcription factor") but provide no indication as to their biological role. Thus, although computational approaches provide valuable guides to experimentation, they do not obviate the need to carry out real experiments to determine protein function [see (*55*)].

An attempt to classify yeast proteins according to their function as conservatively predicted by such computer analyses has been carried out by MIPS (*56*). The yeast cell devotes 11% of its proteome to metabolism; 3% to energy production and storage; 3% to DNA replication, repair, and recombination; 7% to transcription; and 6% to translation. A total of 430 proteins are involved in intracellular trafficking or protein targeting, and 250 proteins have structural roles. Nearly 200 transcription factors have been identified (*57*), as well as 250 primary and secondary transporters (*58*). However, these statistics refer only to yeast proteins for which significant homologs were found.

Another approach that is expected to be greatly facilitated by the availability of all yeast protein sequences is two-dimensional gel electrophoretic analysis, which permits the resolution of more than 2000 soluble protein species (*59*). Unfortunately, many of these membrane proteins are not resolved, and the reproducibility of the two-dimensional electrophoretograms from one laboratory to another is still poor. Identifi-

Table 2. Distribution of genes and other sequence elements. Questionable proteins are defined in (*2*). Hypothetical proteins are the difference between all proteins predicted as ORFs and the questionable proteins. Introns include both experimentally verified examples and those predicted by the EXPLORA program. UTR, untranslated but transcribed regions.

Elements	Chromosome number																
	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	XVI	Total
Sequenced length (kb)	230	813	315	1,532	577	270	1,091	563	440	745	667	1,078	924	784	1,091	948	12,068
Nonsequenced identical repeats																	
Name of unit				*ENA2* and *Y′*		tel		*CUP1*				rDNA and Y′					
Length of unit (kb)				4 and 7		<1		2				9 and 7					
Number of units				2 and 2		1		13				±140 and 2					
Length of repeats (kb)				8 and 14		<1		26				1,260 and 14					1,321
Total length (kb)	230	813	315	1,554	577	271	1,091	589	440	745	667	2,352	924	784	1,091	948	13,389
ORFs (*n*)	110	422	172	812	291	135	572	288	231	387	334	547	487	421	569	497	6,275
Questionable proteins (*n*)	3	30	12	65	13	5	57	12	11	29	20	41	30	23	3	36	390
Hypothetical proteins (*n*)	107	392	160	747	278	130	515	276	220	358	314	506	457	398	566	461	5,885
Introns in ORFs (*n*)	4	18	4	30	13	5	15	15	8	13	11	17	19	15	15	18	220
Introns in UTR (*n*)	0	2	0	1	2	0	5	0	0	0	0	3	0	2	0	0	15
Intact Ty*1* (*n*)	1	2	0	6	1	0	4	1	0	2	0	4	4	2	2	4	33
Intact Ty*2* (*n*)	0	1	1	3	1	1	1	0	0	0	0	2	0	1	2	0	13
Intact Ty*3* (*n*)	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	2
Intact Ty*4* (*n*)	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	3
Intact Ty*5* (*n*)	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
tRNA genes (*n*)	2	13	10	27	20	10	36	11	10	24	16	22	21	16	20	17	275
snRNA genes (*n*)	1	1	2	1	2	0	3	1	1	4	1	3	8	3	7	2	40

cation of the proteins corresponding to given spots by NH_2-terminal protein sequencing has been slow and only about 200 assignments have been made so far. The availability of the predicted amino acid sequence for all yeast proteins should permit a comprehensive analysis of the proteome with the use of rapid and accurate mass spectrometric techniques, so that it should soon become routine to identify all yeast proteins produced under a given set of physiological conditions, or those that are qualitatively or quantitatively modified as a result of the deletion of a specific gene. A new kind of map will then emerge—that of the direct and indirect interactions among all of the members of the yeast proteome. A complete understanding of life at the molecular level cannot be achieved without such knowledge.

Another consequence of defining the yeast proteome is the uncovering of gene products whose existence was hitherto in doubt. For instance, early views that yeast chromosomes were atypical led to the assertion that yeast does not have an H1 histone. In reality, yeast chromosomes contain the full repertoire of eukaryotic histones, including H1, whose gene was found on chromosome XVI (*60*). Another example is the discovery of a yeast gamma tubulin gene on chromosome XII (*14*); this gene had previously eluded yeast geneticists despite intensive efforts by several research groups.

It has been an article of faith for some time that a full understanding of the yeast proteome is a prerequisite for understanding the more complex human proteome; this has become reality with the availability of the complete yeast genome sequence. Nearly half of the proteins known to be defective in human heritable diseases (*61*) show some amino acid sequence similarity to yeast proteins (*62*). Although it is evident that the human genome will specify many proteins that are not found in the yeast proteome, it is reasonable to suggest that the majority of the yeast proteins have human homologs. If so, these human proteins could be classified on the basis of their structural or functional equivalence to members of the yeast proteome.

Genome evolution. The existence of sets of two or more genes encoding proteins with identical or very similar sequences (redundancy) provides the raw material for the evolution of novel functions (*63*). Understanding the true nature of redundancy is one of the major challenges in the quest to elucidate the biological role of every gene in the *S. cerevisiae* genome (*55*). An analysis of the complete genome sequence suggests that it may have undergone duplication events at some point in its evolutionary history. The evidence for such duplications is most readily seen in the pericentric regions and in the central portions of a number of chromosome arms. Although redundancy does occur close to the ends of chromosomes (indeed, the subtelomeric regions are a major repository of redundant sequences), exchanges between such regions are probably too frequent and too recent to help us discern the overall history of the yeast genome.

In *S. cerevisiae*, simple direct repeat clusters (*64*) take several forms, the most typical being dispersed families with related but nonidentical genes scattered singly over many chromosomes. The largest such family comprises the 23 *PAU* genes, which specify the so-called seripauperines (*65*), a set of almost identical serine-poor proteins of unknown function whose ORFs show a very high codon bias and an NH_2-terminal signal sequence (*66*). The *PAU* genes, like the sugar fermentation genes discussed above, reside in the subtelomeric regions. Other dispersed gene families show no obvious chromosomal positioning (such as the 15 members of the *PMT* and *KRE2* families, which encode enzymes involved in the mannosylation of cell wall proteins). Clustered gene families are less common, but a large family of this type occurs on chromosome I, where six related but nonidentical ORFs (YAR023 through YAR033) specify membrane proteins of unknown function (*5*). There are 16 members of this family on six chromosomes. Some are clustered (YHL042 through YHL046 on chromosome VIII), others are scattered singly (YCR007 on chromosome III), and still others are located in subtelomeric regions (YBR302 on chromosome II, YKL219 on chromosome XI, and YHL048 on chromosome VIII). Functional analysis of such a complex family poses a challenge but one that remains within the capability of yeast gene disruption technology.

An analysis of the numerous cluster homology regions (CHRs) revealed by the yeast genome sequence has led to a better understanding of genome evolution. CHRs are large regions in which homologous genes are arranged in the same order, with the same relative transcriptional orientations, on two or more chromosomes. Early reports of CHRs involved a 7.5-kb region on chromosomes V and X (*67*) and a 15-kb region from chromosomes XIV and III (*68*). The latter contains four ORFs that have similarly ordered homologs in the centromeric regions of both chromosomes. One homologous pair consists of two genes, each of which encodes citrate synthase. However, one (*CIT2* on chromosome III) encodes the peroxisomal enzyme, whereas the other (*CIT1* on chromosome XIV) specifies the mitochondrial enzyme. This is probably a good example of evolution through gene duplication, but the situation is even more complicated as a third citrate synthase gene (*CIT3*) has been discovered on chromosome XVI (*68, 69*). Chromosomes IV and II share the longest CHR, comprising a pair of pericentric regions of 120 kb and 170 kb, respectively, that share 18 pairs of homologous genes (13 ORFs and five tRNA genes). The genome has continued to evolve since this ancient duplication occurred: The insertion or deletion of genes has occurred, Ty elements and introns have been lost and gained between the two sets of sequences, and pseudogenes have been generated. In all, at least 10 CHRs (shared with chromosomes II, V, VIII, XII, and XIII) can be recognized on chromosome IV. None of them is found in the central region of the chromosome, which, on the other hand, contains most (7 out of 9) of chromosome IV's complement of Ty elements (Table 2); these may be the cause of the genetic plasticity of this region.

The example of the citrate synthase genes suggests that much of the redundancy in the yeast genome may be more apparent than real. In this case, it was our knowledge of the rules of protein targeting in yeast that allowed us to discern that these genes play different physiological roles. It is likely that a large number of apparently redundant yeast genes are required to deal with physiological challenges that are not encountered in the laboratory environment but that yeast commonly encounters in its natural habitat of the rotting fig (*70*) or grape. Our ability to imagine these conditions and recreate them in the laboratory is severely compromised by our lack of knowledge of the ecology or natural history of *S. cerevisiae*. Indeed, only very recently has it been clearly established that *S. cerevisiae* is found on the surface of the grapes used to make wine (*71*).

What's Next?

For yeast research. New graduate students are already wondering how we all managed in the "dark ages" before the sequence was completed. We must now tackle a much larger challenge, that of elucidating the function of all of the novel genes revealed by that sequence. As with the sequencing project itself, functional analysis will require a worldwide effort. In Europe, a new research network called EUROFAN [for European Functional Analysis Network (*72*)] has been established to undertake the systematic analysis of the function of novel yeast genes. Parallel activities are underway in Germany, Canada, and Japan. In the United States, the National Institutes of Health has recently sent out a request for

applications for "Large-Scale Functional Analysis of the Yeast Genome." Clearly, the yeast research community is mobilizing for the next phase of the campaign to understand how a simple eukaryotic cell works.

In all of this, a common approach is emerging for the deletion of individual genes by a polymerase chain reaction (PCR)–mediated gene replacement technique (*18, 19*). This approach, which relies on the great efficiency and accuracy of mitotic recombination in *S. cerevisiae*, results in the precise deletion of the entire gene and is economical enough to enable producton of the complete set of 6000 single-deletion mutants. This would be a major resource for the scientific community, not only for the functional analysis of the yeast genome itself but also in permitting "functional mapping" of the genomes of higher organisms onto that of yeast. Because of the redundancy problem, and to enable the study of gene interactions, it will also be necessary to construct multiply deleted strains; easy methods to achieve this are already in hand (*73, 74*). All these approaches should make *S. cerevisiae* the eukaryote of choice for the study of functions common to all eukaryotic cells, by reversing the traditional path of genetic research to one in which the study of the gene (or DNA sequence) leads to an understanding of biological function, rather than a change in function leading to the identification of a gene.

For other genomes. Before the release, on 24 April 1996, of the complete yeast genome sequence, two complete bacterial genomes had been made public: the 1.8-Mb sequence of *Haemophilus influenzae* (*75*) and the 0.6-Mb sequence of *Mycoplasma genitalium* (*76*). Another prokaryotic genome, that of *Methanococcus jannaschii* (1.7 Mb), was subsequently released (*77*). The sequences of several other bacterial genomes [*Helicobacter pylori*, *Methanobacterium thermoautotrophicum* (1.7 Mb), *Mycoplasma pneumoniae* (0.8 Mb), and *Synechocystis* sp. (3.6 Mb)] have apparently been completed but were not publicly available at the time this paper went to press. The sequence of at least two dozen other prokaryotic genomes (mostly extremophiles with genome sizes below 2 Mb) is underway. The shotgun sequencing of small bacterial genomes can be completed in less than 6 months at a cost of <$0.50 per base pair (*75, 77*). It is not easy to determine whether such estimates represent full or marginal costs; nevertheless, we can expect many more small genomes to be completed soon. It would be unfortunate if some of these sequences, many of which will be determined through the use of private funds, are not made public in a timely fashion.

For genome sizes between 2 and 6 Mb, sequencing becomes much more complex and expensive. The production of a library comprising a contiguous set of DNA clones (rather than the generation and assembly of the sequence itself) becomes the limiting factor. In the absence of such a library, long-range PCR amplification or direct PCR sequencing, or both, must be employed. The sequencing of genomes larger than 6 Mb typically requires the time-consuming construction of clone libraries in cosmids or other high-capacity vectors and the usually tedious filling-in of the unavoidable, sometimes numerous, and occasionally intractable, gaps in clone coverage. Plans for the determination of medium-sized genome sequences (10 to 100 Mb) almost always underestimate the costs of these essential steps. The existence of two complementary, well-organized, and almost gapless libraries of yeast DNA in cosmid vectors (*78, 79*) was a major factor in the unexpected speed at which the full genome sequence was obtained. After the pilot exercise of chromosome III (*1*), it took only 4 years to complete the remaining 11.8 Mb of the yeast genome. During 1995 alone, more than 6 Mb of final contiguous yeast genomic sequence were obtained. We are confident that it will soon become routine to complete a 10-Mb contig in a year for <$5 million. Nearly complete cosmid libraries of the 15-Mb genome of the fission yeast *S. pombe* are available (*80, 81*), allowing its sequencing to proceed rapidly. If financial support is sustained, *S. pombe* should be one of the next two eukaryotes to have their genome sequences completed [along with the nematode *C. elegans* (*22*), probably in 1998].

Now that the complete sequence of a laboratory strain of *S. cerevisiae* has been obtained, the complete genome sequences of other yeasts of industrial or medical importance are within our reach. Such knowledge should considerably accelerate the development of more productive strains and the search for badly needed antifungal drugs. Unfortunately, the sequence of the important human pathogen *Candida albicans* is proceeding slowly, with limited industrial support (*82, 83*). Complete genome sequencing may be unnecessary when a yeast or fungal genome displays considerable synteny (conservation of gene order) with that of *S. cerevisiae*. For instance, recent studies on *Ashbya gossypii* (a filamentous fungus that is a pathogen of cotton plants) have revealed that most of its ORFs show homology to those of *S. cerevisiae* and that at least a quarter of the clones in an *A. gossypii* genomic bank contain pairs or groups of genes in the same order or relative orientation as their *S. cerevisiae* counterparts (*84*). This gives considerable hope for the rapid analysis of the genomes of a large number of medically and economically important fungi through the use of the *S. cerevisiae* genome sequence as a paradigm. However, this optimism is tempered by the lack of apparent synteny between the *S. cerevisiae* and *S. pombe* genomes. This is perhaps not surprising, as the two species probably diverged from a common ancestor some 1000 million years ago (*85*).

The systematic sequencing of larger model genomes, most notably those of the fruit fly *Drosophila melanogaster* and the dicotyledonous plant *Arabidopsis thaliana*, has now begun. It is doubtful whether either of them will be completed in this century, given that <3% of these 100- to 130-Mb genomes has been systematically sequenced so far. While we await the completion of the human genome sequence sometime around the year 2005, there is danger of dispersing sequencing power among too many model genomes. Instead, it may be desirable to direct sequencing capacity toward eukaryotic parasites (such as *Plasmodium falciparum*, *Trypanosoma cruzei*, *Schistosoma mansoni*, and *Leishmania donovani*) that plague millions of people in developing countries. These genomes are only of intermediate size (30 to 300 Mb) and thus are achievable objects for sequence analysis, provided that funding is increased from its present modest levels. On a world scale, the cost-benefit equation for such projects is overwhelmingly positive.

Pride and Productivity

Two contrasting strategies for gathering genome data have emerged, both of which have been applied to sequencing the yeast genome: the "factory" and the "network" approaches. In the former, sequencing was automated as far as possible and was carried out in large sequencing centers by highly specialized scientists and technicians who may never have seen a yeast outside of a bottle of doubly fermented beer. Their daily notebook, put on the World Wide Web, was fully accessible to the scientific community and was progressively corrected and completed when new information became available. In the network approach, by contrast, yeast genome sequencing was performed in small laboratories by scientists and students deeply committed to the study of particular aspects of yeast molecular biology. These scientists had a special interest in the interpretation of the data and made public only verified and (they hoped) final data, using the same standards as for their normal publications.

In practice, all intermediate forms of approach between these two cultural ex-

tremes have been employed in the yeast genome sequencing project. The network system (to the surprise of some) worked very well: 55% of the total genome sequence was determined by a European Network in which a total of 92 laboratories were involved over the course of the project. The other 45% was obtained by five medium- to large-sized sequencing centers. Over the 6 years of its life, the European Network's performance improved steadily. Its 300-kb fragment of the "international" chromosome XVI was sequenced and made publicly available as rapidly as were the fragments from the three other partners. The sequence quality produced by the two approaches was similar: A large fragment of chromosome XII (170 kb) was deliberately sequenced by both a large center and the European Network; both groups had an extremely low error frequency (one to two sequencing errors per 100 kb). It is estimated that an average of three errors per 10 kb remain in the published version of the yeast genome sequence.

There are at least three reasons for the success of the network. The first is the use of modern informatics technology and the Internet to coordinate the acquisition and analysis of data, as well as to support the general management of the network. The second is that several small laboratories became extremely efficient; the most productive reached 200 kb of finished sequence per year using only two or three people and almost no automation. The third (and most important) reason is that an enthusiastic and competitive team spirit was built up among the small sequencers as, month by month, they watched the data accumulate exponentially toward completion of the genome. The general feeling among the network's participants was that their membership conferred considerable benefits on themselves and on the scientific community as whole.

Whether they worked in large centers or small laboratories, most of the 600 or so scientists involved in sequencing the yeast genome share the feeling that the worldwide ties created by this venture are of inestimable value to the future of yeast research. In Europe especially, a corporate spirit has been engendered that will permit the sharing of data and ideas that will be required to meet the challenge of deciphering the functions and interactions of the novel genes. Nevertheless, it is doubtful that in the future genome sequencing will continue to involve many small laboratories. Increasingly, large-scale sequencing will become the province of the sequencing centers, with the small laboratories being enlisted to sort out problem regions where their specialist knowledge of the organism involved may be of assistance. Enthusiasm, determination, and cooperation (forces that are indispensable under pioneering circumstances) drove this enterprise; we expect these forces will continue to propel us through the next phase of the project.

REFERENCES AND NOTES

1. S. G. Oliver *et al.*, *Nature* **357**, 38 (1992).
2. B. Dujon *et al.*, *ibid.* **369**, 371 (1994).
3. M. Johnston *et al.*, *Science* **265**, 2077 (1994).
4. H. Feldmann *et al.*, *EMBO J.* **13**, 5795 (1994).
5. H. Bussey *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 3809 (1995).
6. Y. Murakami *et al.*, *Nature Genet.* **10**, 261 (1995).
7. F. Galibert *et al.*, *EMBO J.* **15**, 2031 (1996).
8. B. Barrell *et al.*, in preparation.
9. B. Barrell *et al.*, in preparation.
10. H. Bussey *et al.*, in preparation.
11. F. Dietrich *et al.*, in preparation.
12. B. Dujon *et al.*, in preparation.
13. C. Jacq *et al.*, in preparation.
14. M. Johnston *et al.*, in preparation.
15. P. Philippsen *et al.*, in preparation.
16. H. Tettelin *et al.*, in preparation.
17. R. J. Rothstein, *Methods Enzymol.* **101**, 202 (1983).
18. A. Baudin, O. Ozier-Kalogeropoulos, A. Denouel, F. Lacroute, C. Cullin, *Nucleic Acids Res.* **21**, 3329 (1993).
19. A. Wach, A. Brachat, R. Pohlmann, P. Philippsen, *Yeast* **10**, 1793 (1994).
20. R. K. Mortimer, C. R. Contopoulou, J. S. King, *ibid.* **8**, 817 (1992).
21. B. Dujon, *Trends Genet.* **12**, 263 (1996).
22. J. Hodgkin, R. H. A. Plasterk, R. H. Waterston, *Science* **270**, 410 (1994).
23. S. Bowman and B. Barrell, http://www.sanger.ac.uk/yeast/pombe.html
24. R. J. Planta, P. M. Goncalves, W. H. Mager, *Biochem. Cell Biol.* **73**, 825 (1996).
25. B. C. Rymond and M. Rosbash, in *The Molecular Biology of the Yeast* Saccharomyces: *Gene Expression*, E. W. Jones, J. R. Pringle, J. R. Broach, Eds. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1992), pp.143–192.
26. G. R. Fink, *Cell* **49**, 5 (1987).
27. P. M. Sharp and A. T. Lloyd, *Nucleic Acids Res.* **21**, 179 (1993).
28. D. Zenvirth *et al.*, *EMBO J.* **11**, 3441 (1992).
29. B. De Massy, V. Rocco, A. Nicolas, *ibid.* **14**, 4589 (1995).
30. D. B. Kaback, H. Y. Steensma, P. de Jonge, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 3694 (1989).
31. P. Hieter, C. Mann, M. Snyder, R. W. Davis, *Cell* **40**, 381 (1985).
32. A. W. Murray, N. P. Schultes, J. W Szostak, *ibid.* **45**, 529 (1986).
33. H. Y. Steensma, J. C. Crowley, D. B. Kaback, *Mol. Cell. Biol.* **7**, 2894 (1987).
34. M. J Charron, E. Read, S. R. Haut, C. A. Michels, *Genetics* **122**, 307 (1989).
35. M. Carlson, J. L. Celenza, F. J. Eng, *Mol. Cell. Biol.* **5**, 410 (1985).
36. G. I. Naumov, E. S. Naumova, E. J. Louis, *Yeast* **11**, 481 (1995).
37. C. Chan and B.-K. Tye, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 6329 (1980).
38. A. J. Link and M. V. Olson, *Genetics* **127**, 681 (1991).
39. D. F. Voytas and J. D. Boeke, *Nature* **358**, 717 (1992).
40. O. Danilevskaya, A. Lofsky, M.-L Pardue, *Mol. Cell. Biol.* **35S**, 83 (1992).
41. H. Biessmann *et al.*, *ibid.* **12**, 3910 (1992).
42. F. Foury, personal communication.
43. E. J. Louis, *Yeast* **11**, 1553 (1995).
44. E. J. Louis and J. E. Haber, *Genetics* **131**, 559 (1992).
45. C. W Grieder, in *The Eukaryotic Genome: Organisation and Regulation*, P. M. A. Broda, S. G. Oliver, P. F. G. Sims, Eds. (Cambridge Univ. Press, Cambridge, 1993), pp. 31–42.
46. B. L. Wicksteed *et al.*, *Yeast* **10**, 39 (1992).
47. A. Eigel and H. Feldmann, *EMBO J.* **1**, 1245 (1982).
48. J. Gafner, E. M. De Robertis, P. Philippsen, *ibid.* **2**, 583 (1983).
49. J. R. Warmington *et al.*, *Nucleic Acids Res.* **14**, 3475 (1986).
50. J. R. Warmington, R. P. Green, C. S. Newlon, S. G. Oliver, *ibid.* **15**, 8963 (1987).
51. J. Hauber, R. Stucka, R. Krieg, H. Feldmann, *ibid.* **16**, 10623 (1988).
52. H. Lochmüller, R. Stucka, H. Feldmann, *Curr. Genet.* **16**, 247 (1989).
53. H. Ji *et al.*, *Cell* **73**, 1007 (1993).
54. M. R. Wilkins *et al.*, *Bio/Technology* **14**, 61 (1996).
55. S. G. Oliver, *Nature* **379**, 597 (1996).
56. H. W. Mewes, personal communication.
57. V. V. Svetlov and T. G. Cooper, *Yeast* **11**, 1439 (1995).
58. B. Nelissen, P. Mordant, J. L. Jonniaux, R. De Wachter, A. Goffeau, *FEBS Lett.* **377**, 32 (1995).
59. S. Sagliocco *et al.*, *Yeast*, in press.
60. S. C. Ushinsky *et al.*, *ibid.*, in press.
61. http://www3.ncbinim.nih.gov/OMIM/
62. D. E. Bassett, M. S. Boguski, P. Hieter, *Nature* **379**, 589 (1996); http://www.ncbi.nlm.gov/XREFdb/
63. S. Ohno, *Evolution Through Gene Duplication* (Allen and Unwin, London, 1970).
64. M. V. Olson, in *The Molecular Biology of the Yeast* Saccharomyces: *Genome Dynamics, Protein Synthesis, and Energetics*, J. R. Broach, J. R. Pringle, E. W. Jones, Eds. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1991), pp.1–39.
65. M. Viswanathan *et al.*, *Gene* **148**, 149 (1994).
66. B. Purnelle and A. Goffeau, *Yeast*, in press.
67. L. Melnick and F. Sherman, *J. Mol. Biol.* **233**, 372 (1993).
68. D. Lalo, S. Stettler, S. Mariotte, P. P. Slonimski, P. Thuriaux, *C. R. Acad. Sci. Paris* **316**, 367 (1993).
69. A.-M. Bécam, Y. Jia, P. P. Slonimski, C. J. Herbert, *Yeast* **11**, S300 (1995).
70. R. K. Mortimer and J. R. Johnston, *Genetics* **113**, 35 (1989).
71. R. K. Mortimer, T. Torok, P. Romano, G. Suzzi, M. Polsinelli, *Yeast* **11**, S574 (1995).
72. S. G. Oliver, *Trends Genet.* **12**, 241 (1996).
73. F. Langle-Rouault and E. Jacobs, *Nucleic Acids Res.* **23**, 3079 (1995).
74. U. Guldener, S. Heck, T. Fiedler, J. Beinhauer, J. H. Hegemann, *ibid.* **24**, 2519 (1996).
75. R. D. Fleischmann *et al.*, *Science* **269**, 496 (1995).
76. C. M. Frazer *et al.*, *ibid.* **270**, 397 (1995).
77. C. Bult *et al.*, *ibid.* **273**, 1058 (1996).
78. L. Riles *et al.*, *Genetics* **134**, 81 (1993).
79. A. Thierry, L. Gaillon, F. Galibert, B. Dujon ,*Yeast* **11**, 121 (1995).
80. J. D. Hoheisel *et al.*, *Cell* **73**, 109 (1993).
81. T. Mizukami *et al.*, *ibid.*, p. 121.
82. W. S. Chu, B. B. Magee, P. T. Magee. *J. Bacteriol.* **175**, 6637 (1993).
83. http://alces.med.umn.edu/Candida.html
84. R. Altman-Jöhl and P. Philippsen, *Molec. Gen. Genet.* **250**, 69 (1996).
85. P. Russell and P. Nurse, *Cell* **45**, 781 (1986).
86. S. Walsh and B. Barrell, *Trends Genet.* **12**, 276 (1996).