

## Genome sequencing: The complete code for a eukaryotic cell

Mark Johnston

### The complete sequencing of the genome of a simple eukaryotic organism – the budding yeast

***Saccharomyces cerevisiae* – is a milestone for biology, and sets the stage for a complete understanding of how a eukaryotic cell functions.**

Address: Department of Genetics Box 8232, Washington University Medical School, 4566 Scott Avenue, St. Louis, Missouri 63110, USA.

Current Biology 1996, Vol 6 No 5:500–503

© Current Biology Ltd ISSN 0960-9822

The belief that research on the budding yeast, *Saccharomyces cerevisiae*, would reveal processes fundamental to all eukaryotic cells drove its establishment as one of the premier ‘model organisms’ of experimental biology. That promise has been fulfilled beyond all expectations over the last 40 to 50 years. The pioneers in the field were initially attracted to yeast by the ease with which it can be genetically manipulated [1]; the subsequent development of ways of manipulating the yeast genome at will [2,3] attracted a further large cadre of investigators in the 1980s. Yeast has become so popular, and work with it so productive, that it was hard for many of us to imagine that it could become any more attractive as an experimental organism. But that is exactly what has happened, with the announcement on 24 April that the *S. cerevisiae* genome has been completely sequenced. This achievement is already attracting even more scientists to yeast, and it ensures that they will be even more productive than their predecessors. Furthermore, it encourages us to pursue even more vigorously the goal that has been implicit from the beginning: the complete understanding of how a eukaryotic cell functions. The attainment of this lofty goal now seems possible.

The yeast genome project — the goal of which was to identify and map all the genes of the organism — began in earnest in the 1950s [4]. It was not until 1985 that it was clear that the yeast nucleus contains 16 chromosomes [5], and 769 genes were genetically mapped on them [6]. The production of sets of mapped fragment clones spanning nearly the entire 14–15 megabase (Mb) yeast genome [7,8] set the stage for the ultimate genome mapping — determination of the complete DNA sequence — which has now been accomplished through an international collaborative effort of many laboratories. More than half of the sequence (55 %) was completed by a large network of more than 70 laboratories in Europe [9–13]. Significant contributions from other groups — including the Sanger Centre in England (17 %), Washington University in St. Louis (15 %)

[14], Stanford University (7 %), McGill University in Montreal (4 %) [15], and RIKEN University in Japan (2 %) [16] — enabled timely completion of the first eukaryotic genome sequence. The well-coordinated international effort has proved to be very efficient, with surprisingly little duplication of effort. The few instances when a sequence was determined by two groups (about 3 % of the total) was either deliberate, to test sequence quality, or the result of unavoidable overlaps between clones. It can only be hoped that the success of the yeast community in working together to achieve an important goal in a timely and efficient manner, while maintaining high scientific standards, will be replicated by those tackling larger genomes.

The project has met high standards. The sequence was considered complete only when it had been determined on both strands, with no gaps. Furthermore, the sequence of each chromosome was determined all the way out to the telomeres (the G<sub>1-3</sub>A repeats at the ends of each chromosome) [17]. We can thus say that every nucleotide of the yeast genome has been sequenced. We believe that this degree of completion is desirable, given the level of interest in the yeast sequence, and the consequent close scrutiny of it that is expected. The accuracy of the sequence is also very high. All groups (but especially the European network) have expended a great deal of effort to ensure sequence accuracy, and it has paid off. For example, a comparison of 172 kilobases (kb) of sequence determined independently in Europe and St. Louis revealed only 14 discrepancies. More than half of these were due to unavoidable strain or clone polymorphisms (G. Volkaert, personal communication), so the estimated error rate is significantly less than one mistake in 10 kb. Other independent estimates of the error rate yielded slightly higher values [12,13]. (It is worth noting that this is ~5–10-fold more accurate than the piecemeal contributions of yeast sequence in the public databases).

What have we learned from the complete sequence of the yeast genome? We now know that the yeast nuclear genome consists of ~12.5 Mb of unique DNA that encodes the ~6000 proteins that constitute this simple eukaryotic cell (exclusive of the 1–2 Mb rDNA repeat cluster, and a few other repeat families, such as the ~30 kb ‘*CUP1*’ repeats). We expected the genes to be closely packed, but having the opportunity to view the sequence of entire chromosomes [10–16] impresses us with the high information content of the sequence: about 70 % of it codes for protein, with a gene approximately every 2 Kb. We know from piecemeal sequencing that only about

Table 1

Yeast proteins categorized as of March 23, 1996.

By functional category		By major localization category		By molecular environment	
152	Transcription factors	539	Nuclear proteins	549	Integral membrane proteins
109	Protein kinases	420	Cytoplasmic proteins	383	DNA-associated proteins
72	Amino-acid metabolism enzymes	274	Mitochondrial proteins	194	Soluble proteins
50	GTPases	110	Plasma membrane proteins	166	Ribosomal proteins
37	Proteases (non-proteosomal)	77	Endoplasmic reticulum proteins	104	RNA-associated proteins
35	Protein phosphatases	51	Cytoskeletal proteins	99	Peripheral membranes
34	Heat-shock proteins	54	Extracellular and cell-wall proteins	24	Actin-associated proteins
34	tRNA synthetases	51	Vacuolar proteins	16	Tubulin-associated proteins
24	Serine-rich protein family	31	Golgi apparatus proteins	44	Protein synthesis factors
21	Conserved ATPase domain	23	Peroxisomal proteins		
20	Cyclins	28	Vesicular, secretory system proteins		
20	ATP-binding cassette (ABC)				
16	Glucose metabolism enzymes	139	Unspecified membrane proteins		
14	Proteasome subunits	3492	Unknown localization		
12	Ubiquitin-conjugating enzymes				
11	GTPase-activating proteins				
8	Guanine nucleotide exchange factors				

This data is updated daily: see <[http://www.proteome.com/YPD\\_contents\\_by\\_category.html](http://www.proteome.com/YPD_contents_by_category.html)>

4.5 % of yeast genes contain introns [18], nearly all of which are small and located at the extreme 5' end of genes [19], and systematic sequencing has not altered this view. Genes are roughly equally distributed on both DNA strands, and they appear to be randomly orientated with respect to each other. The average gene size is ~480 codons, about half as large again as an average gene of the bacterium *Escherichia coli*. The largest protein (L8004.13), is predicted to contain 4911 amino acids, while the shortest, plasma membrane proteolipid 1 (PMP1), is predicted to contain 38 amino acids.

Perhaps the most startling revelation of the complete yeast genome sequence is that less than half of all the genes have previously been identified through genetic and biochemical analysis, despite intensive analysis over many years by the large and highly productive yeast community. As of March 23, 1996, only about 44 % of the complete gene set had been characterized experimentally — 2307 of the 5271 genes tabulated by Jim Garrels and presented in the Yeast Protein Database (<http://www.proteome.com/YPDhome.html>) [20]. Systematic sequencing has thus yielded a wealth of previously unseen genes, the functions of which must be determined if we are to claim to understand how a simple eukaryotic cell functions. The functions of some of the novel gene products can be guessed from their similarities to other proteins: 30 % (899) of the 2964 previously unrecognized genes show some sequence similarity to other, characterized proteins. However, that leaves 39 % of yeast genes (~2300) for which there are no clues as to their function. While increasingly sensitive homology searches may provide insights into the function of some of these genes [21,22], it is clear that a large fraction of predicted yeast proteins will

not give up their secrets easily. The challenge is to figure out what they do.

One of the things that has probably hindered the genetic identification of all yeast genes is the fact that some of them are redundant. Geneticists have always worried about this, but until now have not known how serious a problem it is. Analysis of the first few chromosomes to be sequenced showed that about 5 % of the genes are the result of recent duplications [10–12,14]. If the criteria for classifying genes as redundant is relaxed, the apparent level of redundancy will of course increase, and more gene pairs will be recognized once the entire sequence is analyzed. Some people believe as many as 30 % of all genes are duplicated [23]. Now that we are able to recognize redundant genes, we should be able to address the functions of some of these previously inaccessible genes.

Even though the analysis of the yeast genome sequence is still far from complete, it is clear what kinds of proteins, and roughly how many of each type constitute this simple eukaryotic cell (Table 1). A large number of proteins seem to function in the nucleus: the major class of proteins by subcellular localization reside in the nucleus; the major class by functional category are transcription factors; and a large class by molecular environment are DNA-associated. This is perhaps not surprising, as the organism must devote much of its effort to the replication, maintenance, proper segregation, and regulation of expression of its 6000 genes. We are also not surprised that the major class of proteins according to molecular environment are integral membrane proteins, many of which are transporters [24], as a yeast cell must obtain diverse nutrients from, and communicate with, its environment. It seems curious,

Table 2

Yeast genome databases.	
Database name	URL
SGD: <i>Saccharomyces</i> Genome Database	<a href="http://genome-www.stanford.edu">http://genome-www.stanford.edu</a>
YPD: Yeast Protein Database	<a href="http://www.proteome.com">http://www.proteome.com</a>
MIPS: European network informatics	<a href="http://speedy.mips.biochem.mpg.de/mips/yeast/">http://speedy.mips.biochem.mpg.de/mips/yeast/</a>
XREFdb: Cross-referencing yeast and other genomes	<a href="http://www.ncbi.nlm.nih.gov/XREFdb/">http://www.ncbi.nlm.nih.gov/XREFdb/</a>
WWW Virtual Library: Yeast	<a href="http://genome-www.stanford.edu/VL-yeast.html">http://genome-www.stanford.edu/VL-yeast.html</a>
GeneQuiz: Comprehensive homology search results	<a href="http://genecrunch.sgi.com/">http://genecrunch.sgi.com/</a>

however, that an organism that can live without carrying out oxidative phosphorylation has a large class of proteins that are localized in the mitochondria.

What should we do with the wealth of data the yeast genome sequence provides? We should, of course, examine it, and fortunately there are several excellent databases that allow us to do so easily (Table 2). These databases, which for the most part are complementary, are sure to facilitate efficient and productive use of the sequence information. The databases are currently being productively used by the community of yeast scientists, but their major impact is likely to be in the guidance they will provide to scientists working on other organisms, including *Homo sapiens*, who come across yeast proteins similar to their proteins of interest.

We need to begin to try to find out what the products of each of the ~3300 new genes uncovered by systematic sequencing do. It is not entirely clear how to go about this task. Some kind of systematic analysis will probably be necessary to begin to understand the roles of genes that provide no hint of their function [25]. Such an approach has been recently organized by the Europeans, as the next phase of their successful sequencing network, and other *ad hoc* efforts are underway [26,27]. Many genes will be disrupted, and a standardized battery of tests carried out to determine the effect of gene loss on the organism. This should yield a wealth of information that is expected to illuminate the functions of some of these genes.

It is likely, however, that this enterprise will encounter a significant problem, as it is expected that disruption of a large fraction of yeast genes (about 50%) will not have any obvious effect on the organism, making it difficult to think of further experiments that might reveal their function. Yeast researchers will, nevertheless, undoubtedly alight upon some of these genes in the course of their

ongoing studies, using the variety of powerful techniques available to them — the two-hybrid method for detecting protein–protein interactions, for example. As the complete sequence of the gene will be on hand, the researcher can proceed immediately with further incisive experimentation, suggested by his or her avenue of the research, which may reveal the function of the protein encoded by the gene. We also expect that the function of many of these genes will be illuminated by analysis of their homologues in other ‘model’ organisms, most of which are richer in phenotypes than yeast, which may provide valuable clues to functions of the encoded proteins [28,29].

A complete understanding of how a simple eukaryotic cell functions will require information from many sources. We hope that the recent completion of the genome sequence will bolster the efforts of the yeast community and attract new investigators to our fold, because our next task — understanding how a simple eukaryotic cell functions — is much more difficult than the task we have just completed. We look forward to the challenge.

#### Acknowledgements

I thank André Goffeau for comments on the manuscript, and for having the foresight to initiate the sequencing project, the skill to organize the sequencers, the determination to maintain high standards, and the fortitude to lead us to the finish line. All of us in the yeast community owe him a great debt. I also thank Jim Garrels for maintaining an excellent database and allowing me to include the information here.

#### References

- Hall MN, Linder P: *The Early Days of Yeast Genetics*. Cold Spring Harbor, New York: Cold Spring Harbor Press; 1993.
- Hinnen A, Hicks JB, Fink GR: Transformation of yeast. *Proc Natl Acad Sci USA* 1978, **75**:1929–1933.
- Scherer S, Davis RW: Replacement of chromosome segments with altered DNA sequences constructed *in vitro*. *Proc Natl Acad Sci USA* 1979, **76**:4951–4955.
- Hawthorne DC, Mortimer RK: Chromosome mapping in *Saccharomyces*: centromere-linked genes. *Genetics* 1960, **45**:1085–1110.
- Carle GF, Olson V: An electrophoretic karyotype for yeast. *Proc Natl Acad Sci USA* 1985, **82**:3756–3760.
- Mortimer RK, Schild D: Genetic map of *Saccharomyces cerevisiae*, Edition 10. *Yeast* 1989, **5**:321–403.
- Riles L, Dutchik JE, Baktha A, McCauley BK, Thayer EC, Ledkie MP, Braden VV, Depke JE, Olson MV: Physical maps of the six smallest chromosomes of *Saccharomyces cerevisiae* at a resolution of 2.6 kilobase pairs. *Genetics* 1993, **134**:81–150.
- Thierry A, Gaillon L, Galibert R, Dujon B: Construction of a complete genomic library of *Saccharomyces cerevisiae* and physical mapping of chromosome XI at 3.7 kb resolution. *Yeast* 1995, **11**:121–135.
- Goffeau A, Vassaroti A: The European project for sequencing the yeast genome. *Res Microbiol* 1991, **142**:901–903.
- Oliver SG, van der Aart QJ, Agostoni-Carbone ML, Aigle M, Alberghina L, Alexandraki D, Antoine G, Anwar R, Ballesta JP, Benit P, et al.: The complete DNA sequence of yeast chromosome III. *Nature* 1992, **357**:38–46.
- Dujon B, Alexandraki D, Andre B, Ansorge W, Baladron V, Ballesta JP, Banrevi A, Bolle PA, Bolotin-Fukuhara M, Bossier P, et al.: Complete DNA sequence of yeast chromosome XI. *Nature* 1994, **369**:371–378.
- Feldmann H, Aigle M, Aljinovic G, Andre B, Baclet MC, Barthe C, Baur A, Becam AM, Biteau N, Boles E, et al.: Complete DNA sequence of yeast chromosome II. *EMBO J* 1994, **13**:5795–5809.
- Galibert F, Alexandraki A, Baur A, Boles E, Chalwatzis N, Chuat J-C, Coster F, Cziepluch C, De Haan M, Domdey H, et al.: Complete nucleotide sequence of *Saccharomyces cerevisiae* chromosome X. *EMBO J* 1996, **15**:2031–2049.

14. Johnston M, Andrews S, Brinkman R, Cooper J, Ding H, Dover J, Du Z, Favello A, Fulton L, Gattung S, *et al.*: **Complete nucleotide sequence of *Saccharomyces cerevisiae* chromosome VIII.** *Science* 1994, **265**:2077–2082.
15. Bussey H, Kaback DB, Zhong W, Vo DT, Clark MW, Fortin N, Hall J, Ouellette BF, Keng T, Barton AB, *et al.*: **The nucleotide sequence of chromosome I from *Saccharomyces cerevisiae*.** *Proc Natl Acad Sci USA* 1995, **92**:3809–3813.
16. Murakami Y, Naitou M, Hagiwara H, Shibata T, Ozawa M, Sasanuma S, Sasanuma M, Tsuchiya Y, Soeda E, Yokoyama K, *et al.*: **Analysis of the nucleotide sequence of chromosome VI of *Saccharomyces cerevisiae*.** *Nat Genet* 1995, **10**:261–268.
17. Louis E, Borts R: **A complete set of marked telomeres in *Saccharomyces cerevisiae* for physical mapping and cloning.** *Genetics* 1995, **139**:125–136.
18. Rodríguez-Medina JR, Rymond BC: **Prevalence and distribution of introns in non-ribosomal protein genes of yeast.** *Mol Gen Genet* 1994, **243**:532–539.
19. Fink GR: **Pseudogenes in yeast?** *Cell* 1987, **49**:5–6.
20. Garrels JI: **YPD – A database for the proteins of *Saccharomyces cerevisiae*.** *Nucleic Acids Res* 1996, **24**:46–49.
21. Koonin EV, Bork P, Sander C: **Yeast chromosome III: new gene functions.** *EMBO J* 1994, **13**:493–503.
22. Ouzounis C, Bork P, Casari G, Sander C: **New protein functions in yeast chromosome VIII.** *Prot Sci* 1995, **4**:2424–2428.
23. Richard G-F, Fairhead C, Dujon B: **Complete transcriptional map of yeast chromosome XI in different life conditions.** *J Mol Biol*, in press.
24. Nelissen B, Mordant P, Jonniaux JL, De Wachter R, Goffeau A: **Phylogenetic classification of the major superfamily of membrane transport facilitators, as deduced from yeast genome sequencing.** *FEBS Lett* 1995, **377**:232–236.
25. Oliver SG: **From DNA sequence to biological function.** *Nature* 1996, **379**:597–600.
26. Burns N, Grimwade B, Ross-Macdonale PB, Choi EY, Finberg K, Roeder GS, Snyder M: **Large-scale analysis of gene expression, protein localization, and gene disruption in *Saccharomyces cerevisiae*.** *Genes Dev* 1994, **8**:1087–1105.
27. Smith V, Botstein D, Brown PO: **Fenetic footprinting: a genomic strategy for determining a gene's function given its sequence.** *Proc Natl Acad Sci USA* 1995, **92**:6479–6483.
28. Tugenreich S, Boguski MS, Seldin MS, Hieter P: **Linking yeast genetics to mammalian genomes: identification and mapping of the human homolog of CDC27 via the expressed sequence tag (EST) data base.** *Proc Natl Acad Sci USA* 1993, **90**:10031–10035.
29. Tugenreich S, Bassett Jr. DE, McKusick VA, Boguski MS, Hieter P: **Genes conserved in yeast and humans.** *Hum Mol Genet* 1994, **3**:1509–1517.