SGM SPECIAL LECTURE

1996 Kathleen Barton-Wright Memorial Lecture

(Delivered at the 136th Meeting of the Society for General Microbiology, 8 January 1997)

Yeast as a navigational aid in genome analysis

Stephen G. Oliver

Tel: +44 161 200 4203. Fax: +44 161 236 0409.
e-mail: steve.oliver@umist.ac.uk

Department of Biochemistry & Applied Molecular Biology, UMIST, PO Box 88, Sackville Street, Manchester M60 1QD, UK

Keywords: *Saccharomyces cerevisiae*, genome analysis, Yeast Genome Project, functional analysis, DNA sequencing

In April 1996, the complete genome sequence of a microbial eukaryote, the brewers' and bakers' yeast *Saccharomyces cerevisiae*, was deposited in the public data libraries (Goffeau *et al*., 1996). This was a landmark event in genetics for two reasons: it was the first complete eukaryotic genome to be sequenced and it was the first genome sequence for an organism that represents a tractable experimental system with a large constituency of active researchers around the globe ready and able to exploit the sequence data. Thus the completion of the yeast genome sequence represented only an important staging post in the Yeast Genome Project; the next, and more difficult, step is the analysis of the functions specified by the 6000 or so protein-encoding genes which the sequence reveals (Oliver, 1996a). This is a large and complex problem but it is, at least, a finite one. Its solution requires us to reverse the normal course of genetical research. Classically, genetics starts by analysing some heritable change in the phenotype of an organism and, by studying the pattern of inheritance of that change, defines the gene that determines it. In recent years, this has been followed by the isolation and sequencing of the defined gene. Thus classical genetics is a 'function first' approach: it proceeds from biological function to DNA sequence. Systematic genome sequencing, in contrast, defines genes *de novo* and requires us to move forward from the gene sequences to the biological functions that they specify. Before I consider how we may attempt to work this trick of performing genetics in the reverse direction, I wish to consider a number of features of genome organization that the complete *S. cerevisiae* sequence has revealed since some of these are central to the problem of functional analysis.

The first yeast chromosome to be sequenced (and, indeed, the first chromosome from any organism) was chromosome III (Oliver *et al*., 1992). When the sequencing was started there were some 37 genetic markers on the map of chromosome III (Mortimer *et al*., 1992), whereas the complete sequence revealed the presence of some 170 genes specifying proteins of $\geqslant 100$ amino acids. It may be contended that the availability of complete DNA sequences for individual chromosomes or entire genomes means that classical genetic maps are now redundant. This is not so; the classical genetic map (obtained, in the case of yeast, mainly by meiotic tetrad analysis) is what it has always been: a measure of the variation in recombination frequency along a chromosome. The difference is that, today, with complete DNA sequences of chromosomes available, we can relate these variations in recombination frequency to the actual physical distance (in base pairs of DNA duplex) between the genetic markers on the chromosome. When this relationship was plotted out for yeast chromosome III (Oliver *et al*., 1992), it revealed an apparently meaningful pattern. Recombination frequency was very low close to the centromere, the genetic map then expanded about half way down each arm before contracting again as the chromosome ends (telomeres) were approached. This pattern of map expansion and contraction has now been shown to be paralleled by the frequency with which double-strand breaks (the structures thought to initiate meiotic recombination in yeast; De Massy *et al*., 1995) occur along the chromosome during meiosis (Zenvirth *et al*., 1992; A. Nicolas, personal communication).

It was of some interest to geneticists to see if a similar relationship existed between the genetic and physical maps obtained for chromosome XI, the second chromosome to be sequenced (Dujon *et al*., 1994). The results were sadly disappointing since there turned out to be major conflicts between the classical genetic map and the physical map derived from the sequence. These conflicts were particularly large on the left arm of the chromosome in which there was a major inversion between the two maps. The conflicts were later found to be due to the accumulation of a number of errors in the genetic map, which has now been shown to be entirely congruent with the physical map derived from the sequence (Simchen *et al*., 1994). However, the sequences of these two chromosomes themselves reveal some higher order similarity between them.

Chromosome III was shown by Sharp & Lloyd (1993) to exhibit a periodic variation in its base composition:

high-AT regions at the centromere and the two chromosome ends being separated by high-GC peaks in the middle of each chromosome arm. The GC content of the chromosome thus parallels the variation in recombination frequency and the frequency of incidence of meiotic double-strand breaks. Chromosome XI (Dujon *et al*., 1994) was found to contain similar 'GC waves', the centromere region again being AT-rich and the period of the oscillations being the same as that found for chromosome III (100 kb for a complete cycle). Most yeast chromosomes exhibit a similar variation in base composition which is found to correlate with variations in gene density along the chromosomes. An exception is chromosome I, where the GC-waves flatten toward the chromosome ends. The 31 kb of DNA at each end of the chromosome are very gene-poor and Bussey *et al*. (1995) have suggested that these terminal domains may act as 'fillers' to increase the size, and hence the stability, of this smallest yeast chromosome. The open reading frames (ORFs) found in these terminal domains are highly repetitive and are frequently interrupted by one or several stop codons.

This apparent redundancy of genetic information is not confined to chromosome I, or the ends of chromosomes, but is found throughout the genome. While some argue that a major expansion of the yeast genome occurred at some time in its evolutionary history (Wolfe & Shields, 1997), there is good evidence for more recent duplication events involving relatively small sections of chromosomes (Melnick & Sherman, 1993; Wicksteed *et al*., 1994). On general evolutionary grounds, one would argue that much of this redundancy is apparent rather than real. Understanding the nature and meaning of genetic redundancy, and developing experimental strategies with which to deal with it, will be central to any systematic attempts to elucidate gene function on a genome-wide scale. One approach, which Dr Ed Louis (Institute of Molecular Medicine, Oxford) and I are adopting, is to try to construct a 'minimalist' yeast genome in which every gene is an essential one. While it must be admitted that the definition of essentiality is (necessarily) an operational one, this exercise will define the basic set of functions required by a eukaryotic cell and enable the contribution of non-essential yeast genes (as well as genes from other organisms) to be assessed by adding them back to this minimalist set.

The quest for the 'minimalist' genome represents a 'top-down' approach to the systematic analysis of gene function. A complementary strategy is to deal with the genes individually. The first, and very important, step in the elucidation of the function of a novel gene is to compare the amino acid sequence of its predicted protein product with those of other protein sequences in the public data libraries to see whether it is similar to a protein of known function that has previously been characterized in another organism. This is an essential first step in functional analysis, but it is not (in itself) sufficient to determine function. First, functionality may be assigned in biochemical terms while giving no clear indication of the biological role of the novel protein. For instance, recognizing that an ORF encodes a protein kinase or phosphatase tells you nothing about the metabolic or developmental pathway in whose regulation such an enzyme may be involved. Second, the assignment of function in the organism where the gene or protein was originally discovered may have been mistaken or, at least, superficial. For instance, yeast chromosome III contains an ORF showing greater than 40% amino acid sequence identity to the NifS proteins of nitrogen-fixing bacteria (Oliver *et al*., 1992). *S. cerevisiae* does not fix N_2, yet the *nifS* homologue is an essential gene. Similar genes have now been found in a number of other bacteria (Leong-Morgenthaler *et al*., 1994; Sun & Setlow, 1993; Mehta & Christen, 1993), none of which fix nitrogen, and experimental and informatic analyses (Leong-Morgenthaler *et al*., 1994; Ouzounis & Sander, 1993) suggest that they encode a class of transaminases that use pyridoxal phosphate as a co-factor. In the nitrogen-fixing bacteria, the *nifS* gene product is now known to be responsible for the insertion of the sulphur into the Fe–S centre of the nitrogenase, again using pyridoxal phosphate as a co-factor (Zheng *et al*., 1993). These examples demonstrate that 'wet' experiments will be necessary to elucidate the functions of the novel genes discovered by systematic sequencing; it cannot be done using the computer alone.

It might be contended that systematic genome sequencing is merely a more efficient way of characterizing genes that will, in any case, be defined by 'classical' or function-first genetics. I do not believe that this is true. First, as the above discussion of recombination and genome organization demonstrated, a complete genome sequence provides a lot more information than just the sequences of its constituent genes. Second, the complete yeast sequence has revealed the presence of a number of genes (e.g. those encoding the snRNP1 protein, Smith & Barrell, 1991; γ-tubulin, Galibert *et al*., 1996; and histone H1, Ushinsky *et al*., 1997) which evaded strenuous attempts at their discovery by more classical ('function-first') routes.

If, instead of the classical route, we are to pursue genetic analysis in the reverse direction, we will need a systematic approach to the elucidation of gene function. This approach should adopt an hierarchical strategy since this will limit the number of experiments to be performed but permit a closer and closer approximation to the function of any individual gene to be achieved. We are adopting such an approach to the functional analysis of the yeast genome in a large European research network, called EUROFAN (Oliver, 1996b).

One of the main aims of the EUROFAN project is to create a library of yeast strains which each carry a specific deletion of an ORF that encodes a protein of unknown function. These deletions are being generated using a PCR-mediated gene replacement protocol developed by Wach and Philippsen (Wach *et al*., 1994; Wach, 1996). In this method, a gene replacement cassette containing a gene (*kanMX*) which confers geneticin resistance on *S. cerevisiae* is tailed, via a PCR reaction,

with sequences homologous to those flanking the target ORF in a yeast chromosome. Geneticin-resistant cells are selected following transformation with this PCR product and 95% or more of the transformants are found to have incorporated *kanMX* in place of the target ORF. The efficiency and accuracy of this replacement event is due to the fact that *kanMX* consists of a drug-resistance determinant from a bacterial transposon which is expressed in yeast by the use of promoter and terminator sequences from the filamentous ascomycete fungus, *Ashbya gossypii*. Thus the only regions of homology to the yeast genome that the replacement cassette contains are those sequences, complementary to the flanks of the target ORF, which the experimenter has designed and used as primers in the PCR reaction. Moreover, competition experiments under a range of physiological conditions (F. Baganz, A. Hayes, D. C. J. Gardner & S. G. Oliver, unpublished) have demonstrated that *kanMX* is a phenotypically neutral marker in yeast growing in the absence of the selective agent, geneticin. Some primary phenotypic analysis of deletion mutants created in this way is performed in the EUROFAN project. In addition, the replacement cassettes, and their cognate genes, are cloned.

These deletants are exploited in EUROFAN by 'Resource Consortia' which carry out tests and analyses capable of application on a genome-wide scale. Relevant genes, together with the appropriate deletant strains and molecular tools, are then passed to specialized 'Functional Analysis Nodes' for more detailed study. Among the Resource Consortia is one dedicated to the quantitative analysis of gene function. The philosophy behind adopting a quantitative approach is that there must be some reason why the growing group of genes that find homologues of unknown function in a number of species have not been found hitherto by molecular genetics with its classical 'function-first' approach. It is possible that these genes have been missed because molecular geneticists usually design their experiments (often with great cleverness) so as to provide qualitative answers to the questions posed. It may be that we are coming to the end of the line as far as this approach is concerned and that there is a group of genes whose functions will only be revealed if we undertake a quantitative analysis of phenotype. The quantitative contribution which these genes make to phenotype may, in some cases, simply be the other side of the coin of redundancy. If a particular gene is a member of a paralogous set of identical (or nearly identical) genes then, provided that they are all regulated in a similar manner and their products are targeted to the same cellular location, then the contribution which an individual member of the set makes to phenotype will, necessarily, be some fraction of the whole.

A conceptual and mathematical framework for the quantitative analysis of gene function is provided by Metabolic Control Analysis (MCA; Kacser & Burns, 1973) in its 'top-down' manifestation (Brand, 1996). Such an approach has just the hierarchical features that I have suggested will be required for a systematic approach to the elucidation of gene function (Oliver, 1996a). At the top level in this hierarchy, quantitative data can most usefully be taken in two main types of experiment. In the first, the effect of deleting or overexpressing a particular gene on the growth rate (or fitness) of the organism is measured. Since the effects are likely to be small, differences in growth rate are most easily revealed in competition experiments since these are very sensitive to such small effects (Danhash *et al*., 1991). Competition experiments using mutant strains generated by either Ty insertions (Smith *et al*., 1995, 1996) or by the PCR-mediated gene replacement protocol described above (F. Baganz, A. Hayes, D. C. J. Gardner & S. G. Oliver, unpublished) have been performed using various forms of quantitative PCR analysis to determine the proportions of the different competing strains in the yeast cell population. The recent development of methods to give deletion mutants specific oligonucleotide tags (so-called 'molecular bar-codes'; Shoemaker *et al*., 1996) offers the prospect of more efficient quantification by employing hybridization-array technology (Schena *et al*., 1996). This would allow a wider range of selective conditions to be investigated.

The second type of data to be generated at the top level of an MCA analysis of gene function is that of the concentration of metabolic intermediates. Such an approach requires a fast and reliable way of sampling the concentration of as many metabolites as possible to produce a kind of 'metabolic snapshot' of each of the deletants (Teusink *et al*., 1997). In collaboration with the laboratories of Douglas Kell (University College of Wales, Aberystwyth) and Simon Gaskell (Michael Barber Centre for Mass Spectrometry, UMIST) we are developing a two-stage strategy for obtaining such data. In the first (or cladistic) phase, deletants of novel genes are grouped with genes of known function by comparing their IR spectra when grown under a range of conditions. This information then decides the type of 'metabolic snapshot' to be taken in the second (or analytical) phase, using tandem mass spectrometry. The combination of this quantitative information on a mutant's metabolic profile together with its calculated μ_{max} under the same conditions (equivalent to the rate of flux when determined under steady-state conditions) provide two essential pieces of information that should allow the site of action of a novel gene product to be located on the yeast metabolic map.

If, by studying this apparently simple eukaryote *S. cerevisiae*, we can work this trick of doing genetics in the reverse direction, and thus obtain at least a good approximation to the function of each of its 6000 genes, I think we will have provided an important 'navigational aid' with which to steer our way through much larger and more complex genomes. A number of positionally cloned human genes that represent the determinants of important single-locus genetic diseases have structural homologues in *S. cerevisiae* (Tugendreich *et al*., 1994). Many of these homologues in yeast were first identified by classical 'function-first' genetics, but others were

revealed by the genome sequencing project (Bassett *et al*., 1996; Oliver, 1996a). In some instances, it has been shown that cDNA copies of the human genes are capable of complementing lesions in the corresponding yeast gene (Ballester *et al*., 1990). We can use our bank of specific deletion mutants of yeast, in a similar way, to map the functionality of the genomes of higher organisms on to that of *S. cerevisiae*. Moreover, we can also use the two-hybrid system (Chien *et al*., 1991) to identify interactions between the proteins of higher organisms and those of yeast. While it is obvious that there will not be a simple one-to-one correspondence between yeast gene functions and those of higher organisms, the correspondences identified should enable the first taxa to be established in an hierarchical taxonomy of gene function in eukaryotes.

Acknowledgements

Work on yeast genome analysis in my laboratory has been supported by the European Commission (in both the Yeast Genome Sequencing and EUROFAN Networks), the BBSRC, the Wellcome Trust, and by Pfizer Central Research, Applied Biosystems, Amersham International and Zeneca.

Bas Teusink is thanked for his critical reading of this manuscript.

References

Ballester, R., Marchuk, D., Boguski, M., Saulino, A., Letcher, R., Wigler, M. & Collins, F. (1990). The *NF*1 locus encodes a protein functionally related to mammalian gap and yeast IRA proteins. *Cell* **63**, 851–859.

Bassett, D. E., Jr, Boguski, M. S. & Hieter, P. (1996). Yeast genes and human disease. *Nature* **379**, 589–590.

Brand, M. D. (1996). Top down metabolic control analysis. *J Theor Biol* **182**, 351–360.

Bussey, H., Kaback, D. B., Zhong, W. W. & 10 other authors (1995). The nucleotide sequence of chromosome I from *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA* **92**, 3809–3813.

Chien, C. T., Bartel, P. L., Sternglanz, R. & Fields, S. (1991). The 2-hybrid system – a method to identify and clone genes for proteins that interact with a protein of interest. *Proc Natl Acad Sci USA* **88**, 9578–9582 .

Danhash, N., Gardner, D. C. J. & Oliver, S. G. (1991). Heritable damage to yeast caused by transformation. *Bio/Technology* **9**, 179–182.

De Massy, B., Rocco, V. & Nicolas, A. (1995). The nucleotide mapping of DNA double-strand breaks at the *CYS3* initiation site of meiotic recombination in *Saccharomyces cerevisiae*. *EMBO J* **14**, 4589–4598 .

Dujon, B., Alexandraki, D., André, B. & 105 other authors (1994). The complete DNA sequence of chromosome XI of *Saccharomyces cerevisiae*. *Nature* **369**, 371–378.

Galibert, Alexandraki, D., Baur, A. & 54 other authors (1996). Complete nucleotide sequence of *Saccharomyces cerevisiae* chromosome X. *EMBO J* **13**, 5795–5809.

Goffeau, A., Barrell, B. G., Bussey, H. & 13 other authors (1996). Life with 6000 genes. *Science* **274**, 546–567.

Kacser, H. & Burns, J. A. (1973). The control of flux. *Symp Soc Exptl Biol* **32**, 65–104.

Leong-Morgenthaler, P., Oliver, S. G., Hottinger, H. & Soll, D. (1994). A *Lactobacillus nif*S-like gene suppresses an *Escherichia coli* transaminase B mutation. *Biochimie* **76**, 45–49.

Mehta, P. & Christen, P. (1993). Homology of pyridoxal-5′-phosphate-dependent aminotransferases with the *cobC* (cobalamin synthesis), *nifS* (nitrogen fixation), *pabC* (*p*-amino-benzoate synthesis) and *malY* (abolishing endogenous induction of the maltose system) gene products. *Eur J Biochem* **211**, 373–376.

Melnick, L. & Sherman, F. (1993). The gene clusters *ARC* and *COR* on chromosome V and chromosome X, respectively, of *Saccharomyces cerevisiae* share a common ancestry. *J Mol Biol* **233**, 372–388.

Mortimer, R. K., Contopoulou, C. R. & King, J. S. (1992). Genetic and physical maps of *Saccharomyces cerevisiae*, Edition 11. *Yeast* **8**, 817–902.

Oliver, S. G. (1996a). From DNA sequence to biological function. *Nature* **379**, 597–600.

Oliver, S. G. (1996b). A network approach to the systematic analysis of yeast gene function. *Trends Genet* **12**, 241–242.

Oliver, S. G., van der Aart, Q. J. M., Agostoni-Carbone, M. L. & 144 other authors (1992). The complete DNA sequence of yeast chromosome III. *Nature* **357**, 38–46.

Ouzounis, C. & Sander, C. (1993). Homology of the NifS family of proteins to a new class of pyridoxal phosphate-dependent enzymes. *FEBS Lett* **322**, 159–164 .

Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O. & Davis, R. W. (1996). Parallel human genome analysis – microarray-based expression monitoring of 1000 genes. *Proc Natl Acad Sci USA* **93**, 10614–10619.

Sharp, P. M. & Lloyd, A. T. (1993). Regional base composition variation along yeast chromosome III – evolution of chromosome primary structure. *Nucleic Acids Res* **21**, 179–183.

Shoemaker, D. D., Lashkari, D. A., Morris, D., Mittmann, M. & Davis, R. W. (1996). Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nat Genet* **14**, 450–456.

Simchen, G., Chapman, K. B., Caputo, E., Nam, K., Riles, L., Levin, D. E. & Boeke, J. D. (1994). Mapping of *dbr1* and *ypk1* suggests a major revision of the genetic map of the left arm of *Saccharomyces cerevisiae* chromosome XI. *Genetics* **138**, 283–287.

Smith, V. & Barrell, B. G. (1991). Cloning of a yeast U1 snRNP 70k protein homolog – functional conservation of an RNA-binding domain between humans and yeast. *EMBO J* **10**, 2627–2634.

Smith, V., Botstein, D. & Brown, P. O. (1995). Genetic footprinting – a genomic strategy for determining a gene's function given its sequence. *Proc Nat Acad Sci USA* **92**, 6479–6483.

Smith, V., Chou, K. N., Lashkari, D., Botstein, D. & Brown, P. O. (1996). Functional analysis of the genes of yeast chromosome V by genetic footprinting. *Science* **274**, 2069–2074.

Sun, D. & Setlow, P. (1993). Cloning and nucleotide sequence of the *Bacillus subtilis nad*B gene and a *nif*S-like gene both of which are essential for NAD biosynthesis. *J Bacteriol* **175**, 1423–1432.

Teusink, B., Baganz, F., Westerhoff, H. V. & Oliver, S. G. (1997). Metabolic Control Analysis as a tool in the elucidation of the function of novel genes. *Methods Microbiol* (in press).

Tugendreich, S., Bassett, D. E., Jr, McKusick, V. A., Boguski, M. S. & Hieter, P. (1994). Genes conserved in yeast and humans. *Hum Mol Genet* **3**, 1509–1517.

Ushinsky, S., Bussey, H., Ahmed, A. A., Wang, Y., Friesen, J., Williams, B. A. & Storms, R. K. (1997). Histone H1 in *S. cerevisiae*.*Yeast* **13**, 151–161.

Wach, A. (1996). PCR-synthesis of marker cassettes with long flanking homology regions for gene disruptions in *Saccharomyces cerevisiae*. *Yeast* **12**, 259–265.

Wach, A., Brachat, A., Pohlmann, R. & Philippsen, P. (1994). New heterologous modules for classical or PCR-based gene disruptions in *Saccharomyces cerevisiae*. *Yeast* **10**, 1793–1808.

Wicksteed, B. L., Collins, I., Dershowitz, A., Stateva, L. I., Green, R. P., Oliver, S. G., Brown, A. J. P. & Newlon, C. S. (1994). A physical comparison of chromosome III in six strains of *Saccharomyces cerevisiae*. *Yeast* **10**, 39–47.

Wolfe, K. & Shields, D. (1997). Yeast gene duplications. *http://acer.gen.tcd.ie/ijhwolfe/yeast/nova/index.html*

Zenvirth, D., Arbel, T., Sherman, A., Goldway, M., Klein, S. & Simchen, G. (1992). Multiple sites for double-strand breaks in whole meiotic chromosomes of *Saccharomyces cerevisiae*. *EMBO J* **11**, 3441–3447 .

Zheng, L., White, R. H., Cash, V. L., Jack, R. F. & Dean, D. R. (1993). Cysteine desulfurase activity indicates a role for NifS I metallocluster biosynthesis. *Proc Natl Acad Sci USA* **90**, 2754–2758.