

The yeast genome: on the road to the Golden Age

Mark Johnston

Having the complete genome sequence of *Saccharomyces cerevisiae* makes us aware of the ultimate goal of yeast molecular biology: the 'solution' of the cell, that is, an understanding of the function of all ~6000 proteins (and a few RNAs) and how they interact with each other and the environment. The recent development of 'genomic' approaches for studying gene function makes this goal seem reachable in the foreseeable future. When this is accomplished, we will have entered a Golden Age, when we will have the information necessary for designing truly incisive experiments to reveal biological function.

Addresses

Department of Genetics, Box 8232, Washington University School of Medicine, 660 Euclid Avenue, St Louis, Missouri 63113, USA:
e-mail: mj@genetics.wustl.edu

Current Opinion in Genetics & Development 2000, **10**:617–623

0959-437X/00/\$ – see front matter
© 2000 Elsevier Science Ltd. All rights reserved.

Abbreviations

ORF open reading frame
SmORF small ORF
YKO yeast gene knockout (collection)

Introduction

Unlike an earnest tourist, who knows exactly where he is going and when he will arrive, a scientist usually has only a vague idea of his destination, and seldom comes even close to predicting his arrival time. When Bob Mortimer embarked in the 1950s on the scientific journey that was the yeast genome project [1], he could not have foreseen that his genetic map would ultimately be at single nucleotide resolution, and I doubt he had even an inkling that it might lead to the identification of *all* the genes that comprise this relatively simple eukaryotic cell. Surely he did not imagine these things would happen before the end of the century. Twenty years later, the goal of determining the DNA sequence of the yeast genome was certainly in Maynard Olson's sights when he initiated its physical mapping in the late 1970s [2] but I doubt he saw the destination as understanding the function of all of the organism's genes. By the time DNA sequencing had begun in the late 1980s [3], the goal of identifying all yeast genes was explicit, but no one came close to predicting when the job would be done (*Science* 1992, 256:462). Furthermore, it seems unlikely that anyone imagined at that time that the contributions each of the genes makes to the life of the cell might be known in the foreseeable future. Now that we have had over four years to develop and implement new experimental approaches for exploiting the complete yeast genome sequence, our destination is clear: the 'solution' of the organism, that is, an understanding of the functions and inter-relationships of each of its ~6000 genes. When will we reach this goal and what will it mean when we do?

The yeast genome: features

Global features of the genome

When the first complete DNA sequence of a eukaryotic genome became available in 1996 [4], it was thought that long-range features of yeast genome architecture might be apparent. But early excitement over seemingly periodic variations in G+C content of some yeast chromosomes waned as it became clear that this was not the case throughout the genome [5]. Although the nucleotide composition within and between chromosomes clearly varies [6,7], the biological significance (if any) of this observation is unclear. Unclear also is the significance of the non-random distribution of microsatellite sequences [8–10].

Perhaps the most significant insight into genome organization is the realization that the entire genome probably duplicated ~150 million years ago [11]. Pieces of the duplicated chromosomes subsequently translocated to different chromosomes, followed by loss of most (~92%) of the genes from one or the other duplicated copy, resulting in the 55 different stretches of duplicated genes apparent in the sequence today [12,13]. A key observation that supports this idea is that the duplicated genes in each block are almost always in the same order and relative orientation, and their order is similar to that inferred to exist prior to genome duplication [14,15].

Finding all the genes

Even the seemingly simple, relatively prosaic goal of the yeast genome project — to identify all the genes of the organism — has not been met. The genome contains ~6300 potential genes — the MIPS database [16] lists 6340 open reading frames (ORFs); SGD [17] lists 6281; YPD [18] lists 6142 proteins — but it will be a long time until all yeast genes have been identified unequivocally. One reason is that small genes (<100 codons) are difficult to recognize and too numerous to annotate [19]. About 160 of these small ORFs — smORFs, also called nORFs for 'non-annotated ORFs' — appear to be transcribed [20], suggesting that they are actual genes. A recent screen for expressed regions of the yeast genome using a transposon containing a reporter of gene expression has identified 328 expressed smORFs [21]. Even these kinds of evidence, however, do not definitively identify a smORF as an actual gene as background transcription/translation of the genome might be what is being detected in some cases. In fact, ~15% of the reporter gene transposons that landed in smORFs are within a larger annotated ORF, in the same transcriptional orientation but in a different translational reading frame. Reporter gene expression in these cases is probably caused by frameshifting of translation of the larger ORF, which seems to occur surprisingly frequently [21]. One wonders how many of the other 85% of the smORFs identified in this way are being expressed

fortuitously. In any case, a reasonable guess is that ~250 smORFs (\pm ~50) are real genes.

Many of the annotated ORFs (those \geq 100 codons) are undoubtedly not genes, but how many? Approximately 10% of them can be expected in a random sequence [22], implying that ~600 annotated ORFs do not encode proteins, giving yeast ~6000 genes (~6300 ORFs minus ~600 false ORFs plus ~250 smORFs). Some people believe that even more of the ORFs are imposters because many have a nucleotide composition that is unlike that of true genes. If correct, this would leave yeast with only ~4800 genes [23,24]. It would be surprising if a yeast cell can be built with about the same number of genes as *Escherichia coli*. Indeed, our comparison of the DNA sequences of several *Saccharomyces* species shows that the sequences of most of the ORFs of unknown function are well conserved through evolution, suggesting that most of them really are genes (P Cliften *et al.*, unpublished data). This important issue cannot be resolved until we have functional data for every gene.

Despite the fact that introns in yeast are rare and small, it is still not straightforward to predict how the mRNA of a yeast gene will be spliced. When this was tested for several mRNAs thought to contain introns, ~30% of the predictions turned out to be wrong [25]! (One shudders to think what this means for the error rate in larger eukaryotes!) Data from this and a variety of other studies have been used to improve intron predictions, yielding 231 yeast genes predicted to contain introns, though this is admittedly not a complete set [26].

Some genes, of course, do not encode proteins, making them particularly difficult to recognize. It is fairly easy to recognize tRNAs, thus it is likely that all of them (274) have been found [27,28]. Forty one snoRNAs that guide most of the ribose methylations of rRNA were predicted on the basis of their characteristic sequence elements, then confirmed experimentally [29*]. A few other probable non-protein coding RNAs were discovered by searching the genome for RNA Polymerase III promoter sequences and by looking in large intergenic regions for transcripts [30]. Undoubtedly, many more genes encoding either structural or catalytic RNAs remain to be identified. Some of them probably lie in the several sequences that do not contain ORFs that we have found to be highly conserved in several *Saccharomyces* species (P Cliften *et al.*, unpublished data).

Finding sequence motifs

In addition to identifying all the genes, we need to understand the networks regulating their expression before we can claim to have 'solved' the organism. A big step in this direction would be identification of DNA sequences that most likely regulate gene expression. This is easy to do if the transcription factor and its binding site are known, allowing one to identify all the genes in the genome potentially regulated by that protein (e.g. see [31]),

though the predictions must be experimentally verified (e.g. see [32]). A more challenging problem is finding regulatory sequences when nothing is known about the regulatory mechanism. This seemingly impossible task has been accomplished by several groups that have devised methods for identifying DNA sequence motifs that are over-represented in a set of genes thought to be related (usually because they exhibit a similar pattern of expression, or because they are expected to function in the same cellular pathway) [33–40]. These methods seem to work well because they usually find the correct transcription factor binding sites among genes whose mechanism of regulation is known. Revealed also are other, previously unrecognized sequence motifs that can be tested for their role in regulating gene expression. Similar methods have been used to identify sequences in mRNAs that direct their splicing [41] and polyadenylation [42]. Given the ease with which genes can be placed into likely functional groups according to their pattern of gene expression [37,43], or even their database annotations [34], it is not unreasonable to expect that most (perhaps all?) gene regulatory sequences will be identified soon, providing an entrée to the underlying gene regulatory mechanisms and networks.

The yeast genome: function

Now that most yeast genes have been identified, the challenge is to define the function of each of their products. This is a big job because the portfolio of yeast genes grew >3-fold when the genome sequence was completed [4,5]. The current classification of yeast proteins is shown in Figure 1 (for a more detailed list, see <http://www.proteome.com/databases/YPD/> or <http://www.mips.biochem.mpg.de/proj/yeast/catalogues/fun-cat/index.html>). Note that the major category is 'unknown', so we are far from our goal of 'solving' the organism. To achieve this, we will need to determine the function of the substantial number of unclassified proteins, learn much more about the poorly defined ones, and understand how all of them interact with each other and with the environment. Sometimes (about one-third of the time) we can infer function of a protein on the basis of its similarity to proteins the functions of which have been characterized in other organisms. The clever application of computational tools can also provide hints of protein function by linking a protein to a group of proteins, some of which may be of known function on the basis of several characteristics of the protein, including its tendency to co-evolve with proteins in other organisms (its phylogenetic profile), or to be fused to other proteins in different organisms [44]. Of course, experimental evidence is the ultimate arbiter of gene function, and must be collected for every gene in the genome before we can begin to pretend that the organism is 'solved'. The availability of new resources and approaches for analyzing the entire genome make me optimistic this can be achieved in the foreseeable future. These approaches have the potential of providing for each protein the six fundamental pieces of information needed for predicting its function (described below).

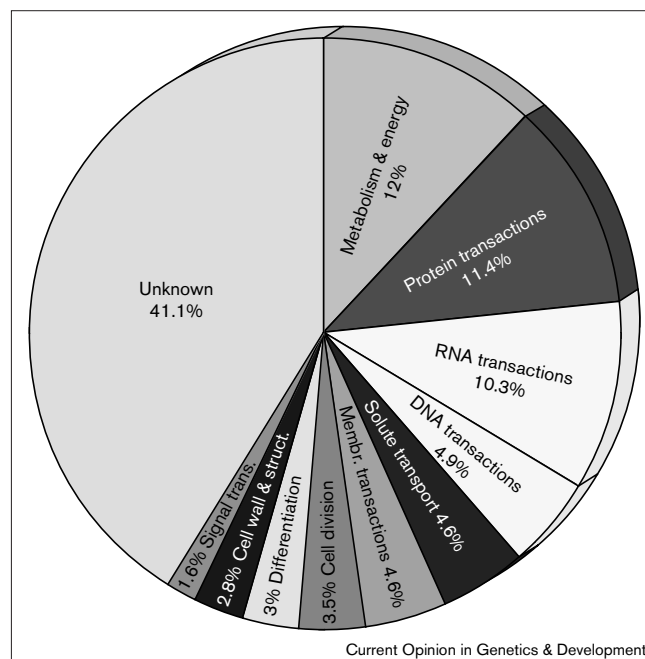
The effect of loss of protein function

Although yeast geneticists no longer need mutations to lead them to genes, mutations in an ORF are still necessary to confirm that it *is* a gene and to provide clues to the function of the product it encodes. Fortunately, we no longer have to rely on fickle fate to provide mutations in genes we are interested in: we can systematically make mutations in each of them — or at least the ones we are able to predict. This has been accomplished recently by an international consortium of labs that produced the yeast gene knockout (YKO) collection, an array of ~6000 mutants, each missing a different gene [45]. The YKO collection is sure to be a useful resource for discovering gene function, as it should enable identification of *all* genes whose loss of function disrupts a cellular process under study. Even redundant genes in a pathway might be identified as it may be possible to use the YKO collection to test the phenotypes of all 36 million double mutants (C Boone, personal communication).

The initial forays into phenotypic screening of all mutants are encouraging. About 33% of the mutants exhibit an obvious growth defect — ~18% are essential, ~15% grow slowly on rich medium [45,46]. When a larger number of diverse phenotypes were screened, as many as two-thirds of the mutants exhibited at least one obvious phenotype [21,47,48]. Although most of the phenotypes provide few clear hints about protein function (e.g. temperature-sensitive growth reveals little about gene function), it is hoped that the accumulation of several phenotypes exhibited by a mutant will enable formulation of hypotheses of protein function. Indeed, recognizing groups of mutants exhibiting similar phenotypes can suggest protein function [21]: if a group contains genes that encode proteins whose function is known, the unknown genes in the group are implicated in that function. Distribution of the YKO collection to many laboratories should result in a plethora of phenotypes for many mutants, and contribute significantly to ‘solving’ the organism, (see http://sequence-www.stanford.edu/group/yeast_deletion_project/ for information on how to obtain the YKO collection).

Phenotyping the ~6000 mutants of the YKO collection can be accelerated using a clever approach that enables testing the fitness of all of them in a single overnight growth experiment. This method relies on the ability to detect each mutant through the unique DNA sequence tag (a ‘molecular bar code’) it carries, which was inserted into its genome during the gene-disruption process [49]). Those mutants that are underrepresented in a culture of all 6000 mutants after growth under a particular condition identify the genes required for growth under that condition (in rare cases, a mutant may be over-represented after growth of the population, which would indicate that the normal gene inhibits growth under that condition). Application of this approach to >2000 mutants yielded a measurable phenotype for ~40% of them [45]. Similar results were obtained by an analogous method that employs random mutagenesis of a

Figure 1



Categories of yeast protein function, according to information in the Yeast Protein Database [18] as of August, 2000. Note that each protein was assigned only one function, despite the fact that many proteins act in more than one process. For a more detailed list of protein function categories see <http://www.proteome.com/databases/YPD/> or <http://www.mips.biochem.mpg.de/proj/yeast/catalogues/funcat/index.html>.

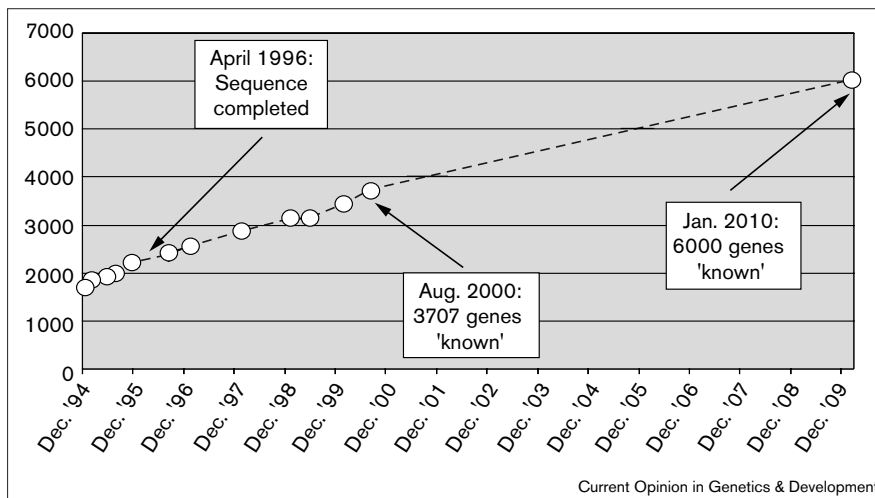
large population of cells [50]. Although these approaches are certainly efficient, I suspect they will contribute little to ‘solving’ the organism because the phenotypes they provide — usually subtle reductions in growth rate — will seldom lead to testable hypotheses of protein function. Nevertheless, the ‘bar codes’ have tremendous potential in other applications [51•] and I expect that additional innovative uses for them will be found.

One especially informative mutant phenotype that has just begun to be exploited is genome-wide changes in gene expression. Loss of function of a gene often causes the expression of many other genes to change. Genes whose inactivation causes similar changes in expression (both in the levels and the timing) of a similar subset of other genes are usually involved in the same cellular process [52••]. Thus, if mutations in several genes cause the same pattern of changes in expression of the same set of genes, the unknown genes in that group are implicated in the process in which the known genes are involved. This approach has recently been used to discover the function of several yeast genes [52••], and promises to greatly accelerate progress toward ‘solving’ the organism.

The effect of protein overexpression

The other side of the phenotypic coin is gene overexpression, which has proven fruitful for identifying genes

Figure 2



Progress in 'characterizing' yeast protein function. The number of yeast proteins listed as 'known' in the Yeast Protein Database [18] on each date tabulated (<http://www.proteome.com/databases/YPD/YPDfacts.html>) is plotted. The dotted line is an extrapolation of the observed rate of progress.

involved in many cellular processes [53]. Although genes that cause a phenotype when expressed at a high level can be selected from random clone libraries (e.g. see [54]), one would like to be able to test all genes systematically for their effect on the cell when present at high levels. Unfortunately, no resource for doing this is currently available.

Regulation of gene expression

Patterns of gene expression can provide clues to protein function because genes whose expression pattern is similar often encode proteins of related function. Thus, if a group of similarly regulated genes includes some whose function is known, the unknown genes in the group are implicated in that function [43,55]. The development of methods to measure expression of all genes of an organism in a single experiment (using 'gene chips' [55,56]), coupled with algorithms to sort the genes into similarly regulated clusters [43], has already provided hints of the function of hundreds of yeast proteins [37,52**]. It is expected that further implementation of this approach will provide clues to the function of many (perhaps most) of the uncharacterized yeast proteins. In addition, whole genome expression measurements have been used to identify all the genes that respond to some transcription factors [57–62] and to dissect several signal transduction pathways [63–66], thereby revealing regulatory networks within the cell. The prospects for 'solving' yeast are greatly improved by this revolutionary technology, the impact of which is just beginning to be realized. Enthusiasm for this approach, however, is tempered by the realization that mRNA and protein levels are often not correlated [67].

Location of the protein in the cell

The location of a protein in the cell can provide an obvious clue to its function. It seems that epitope-tagging of yeast proteins will enable the subcellular location of about a third of them to be determined [21], thereby adding to our list of functional clues for many proteins.

Interacting proteins

Proteins that physically interact usually work in the same process, so if a protein whose function is known is found to physically interact with one of unknown function, the role of the unknown protein is virtually established. Even when a protein's function is known, knowledge of the other proteins it interacts with is necessary to define more completely its role in the cell. A map of all the protein–protein interactions in the yeast cell is being constructed by the two-hybrid method, using arrays of 6000 yeast genes fused to a DNA-binding or transcriptional activation domain [68,69]. Although this approach clearly yields meaningful data [68–70], some of the interactions are almost certainly artifacts, so the results must, of course, be verified by more direct experiments (like all of the experiments described here). A particularly promising method for purifying protein complexes and directly identifying their components has been described recently [71*]. It is clear that these approaches will provide a wealth of valuable clues that are certain to contribute significantly to 'solving' the organism.

Protein activity

Perhaps the most definitive indication of a protein's function is its enzymatic activity. Identifying a protein that has a particular activity is a major challenge because substantial effort is usually required for its purification. The recent development of an array of 6000 yeast strains, each expressing a different epitope-tagged protein, greatly accelerates the identification of protein activity [72**] because each protein can be easily purified in one step using the epitope tag. Pooling of the cells expressing the epitope-tagged proteins increases the efficiency of the process. In principle, only 84 assays are required to identify the protein that possesses the desired activity: 64 pools, each containing 96 different epitope-tagged proteins are first assayed, then the protein in the pool exhibiting the activity is identified by assaying the 20 pools of cells from the 8 rows and from the

12 columns of a 96-well microtiter plate. This reduces the time required for identifying a protein's enzymatic activity from months or years to just a few days! One could also imagine using this array of proteins to identify protein-protein interactions, substrates of protein-modifying enzymes, or drug targets. The impact this approach will have on 'solving' the organism is clearly immense.

The encyclopedia of yeast is sure to expand rapidly over the next several years as these new techniques and approaches for analyzing all the genes of the organism take root in the research community. Fortunately, we have outstanding curators of the encyclopedia [16–18], so we have ready access to this growing body of knowledge.

The yeast genome: future

Like an earnest tourist, we can now clearly see our destination — the 'solution' of the organism. Can we predict accurately when we will arrive there? Our rate of progress in discovering yeast gene function has been remarkably linear (Figure 2), and suggests we will know something informative about each of the 6000 yeast proteins in ~10 years. In fact, this is probably a pessimistic estimate because spread throughout the yeast research community of the productive new approaches for whole-genome analysis described above should greatly accelerate gene function discovery.

Conclusions

When we reach our goal in a few years, will we really have 'solved' the organism? The answer, of course, is an emphatic *No!* Once we have 'characterized' all 6000 yeast proteins, we will still know little about the function of most of them. (Just look in YPD: you are unlikely to be satisfied with our understanding of most of the proteins currently classified as 'characterized'.) Rather than a 'solution' of the organism, what we will soon produce is a fund of knowledge about it that will enable us to devise incisive experiments aimed at truly understanding how a yeast cell is built and functions. In the past — and, arguably, even now — we have been working with a woefully incomplete information infrastructure. The availability of the DNA sequences of whole genomes partially redressed that deficiency; 'characterization' of all 6000 yeast proteins will complete the first edition of the encyclopedia of a eukaryotic cell. But, just as an encyclopedia is only the first resource to which one turns for fundamental (but usually superficial) information on a topic, our 'solution' of yeast will serve as a resource to guide further, more productive experimentation. Completion of the writing of this encyclopedia of a eukaryotic cell will herald the Golden Age of biology, when we can expect to gain a deep and genuine understanding of how cells and organisms function. I am looking forward to being a yeast geneticist in the coming Golden Age!

Acknowledgements

I thank the curators of the yeast genome databases (described in references [16–18]) for their outstanding service to all of us in the yeast community. I'm especially grateful to Mike Cherry at SGD and Peter Hodges at YPD for promptly answering my frequent questions.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
 - of outstanding interest
1. Hawthorne D, Mortimer RK: **Chromosome mapping in *Saccharomyces*: centromere-linked genes.** *Genetics* 1960, **45**:1085-1110.
 2. Olson MV, Dutchik JE, Graham MY, Brodeur GM, Helms C, Frank M, MacCollin M, Scheinman R, Frank T: **Random-clone strategy for genomic restriction mapping in yeast.** *Proc Natl Acad Sci USA* 1986, **83**:7826-7830.
 3. Goffeau A, Vassarotti A: **The European Project for Sequencing the Yeast Genome.** *Fresenius J An Chem* 1990, **337**:29-30.
 4. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M *et al.*: **Life with 6000 genes.** *Science* 1996, **274**:546,563-567.
 5. Dujon B: **The yeast genome project: what did we learn?** *Trends Genet* 1996, **12**:263-270.
 6. Li W, Stolovitzky G, Bernaola-Galvan P, Oliver JL: **Compositional heterogeneity within, and uniformity between, DNA sequences of yeast chromosomes.** *Genome Res* 1998, **8**:916-928.
 7. Bradnam KR, Seoighe C, Sharp PM, Wolfe KH: **G+C content variation along and among *Saccharomyces cerevisiae* chromosomes.** *Mol Biol Evol* 1999, **16**:666-675.
 8. Richard GF, Hennequin C, Thierry A, Dujon B: **Trinucleotide repeats and other microsatellites in yeasts.** *Res Microbiol* 1999, **150**:589-602.
 9. Richard GF, Dujon B: **Trinucleotide repeats in yeast.** *Res Microbiol* 1997, **148**:731-744.
 10. Young ET, Sloan JS, Van Riper K: **Trinucleotide repeats are clustered in regulatory genes in *Saccharomyces cerevisiae*.** *Genetics* 2000, **154**:1053-1068.
 11. Wolfe KH, Shields DC: **Molecular evidence for an ancient duplication of the entire yeast genome.** *Nature* 1997, **387**:708-713.
 12. Seoighe C, Wolfe KH: **Updated map of duplicated regions in the yeast genome.** *Gene* 1999, **238**:253-261.
 13. Seoighe C, Wolfe KH: **Extent of genomic rearrangement after genome duplication in yeast.** *Proc Natl Acad Sci USA* 1998, **95**:4447-4452.
 14. Keogh RS, Seoighe C, Wolfe KH: **Evolution of gene order and chromosome number in *Saccharomyces*, *Kluyveromyces* and related fungi.** *Yeast* 1998, **14**:443-457.
 15. Skrabanek L, Wolfe KH: **Eukaryote genome duplication — where's the evidence?** *Curr Opin Genet Dev* 1998, **8**:694-700.
 16. Mewes HW, Heumann K, Kaps A, Mayer K, Pfeiffer F, Stocker S, Frishman D: **MIPS: a database for genomes and protein sequences.** *Nucleic Acids Res* 1999, **27**:44-48.
 17. Ball CA, Dolinski K, Dwight SS, Harris MA, Issel-Tarver L, Kasarskis A, Scafe CR, Sherlock G, Binkley G, Jin H *et al.*: **Integrating functional genomic information into the *Saccharomyces* genome database.** *Nucleic Acids Res* 2000, **28**:77-80.
 18. Costanzo MC, Hogan JD, Cusick ME, Davis BP, Fancher AM, Hodges PE, Kondu P, Lengieza C, Lew-Smith JE, Lingner C *et al.*: **The yeast proteome database (YPD) and *Caenorhabditis elegans* proteome database (WormPD): comprehensive resources for the organization and comparison of model organism protein information.** *Nucleic Acids Res* 2000, **28**:73-76.
 19. Basrai MA, Hieter P, Boeke JD: **Small open reading frames: beautiful needles in the haystack.** *Genome Res* 1997, **7**:768-771.
 20. Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE Jr, Hieter P, Vogelstein B, Kinzler KW: **Characterization of the yeast transcriptome.** *Cell* 1997, **88**:243-251.
 21. Ross-Macdonald P, Coelho PS, Roemer T, Agarwal S, Kumar A, Jansen R, Cheung KH, Sheehan A, Symoniatis D, Umansky L *et al.*: **Large-scale analysis of the yeast genome by transposon tagging and gene disruption.** *Nature* 1999, **402**:413-418.

22. 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.
23. Kowalczyk M, Mackiewicz P, Gierlik A, Dudek MR, Cebrat S: **Total number of coding open reading frames in the yeast genome.** *Yeast* 1999, **15**:1031-1034.
24. Mackiewicz P, Kowalczyk M, Gierlik A, Dudek MR, Cebrat S: **Origin and properties of non-coding ORFs in the yeast genome.** *Nucleic Acids Res* 1999, **27**:3503-3509.
25. Davis CA, Grate L, Spingola M, Ares M Jr: **Test of intron predictions reveals novel splice sites, alternatively spliced mRNAs and new introns in meiotically regulated genes of yeast.** *Nucleic Acids Res* 2000, **28**:1700-1706.
26. Spingola M, Grate L, Haussler D, Ares M Jr: **Genome-wide bioinformatic and molecular analysis of introns in *Saccharomyces cerevisiae*.** *RNA* 1999, **5**:221-234.
27. Percudani R, Pavasi A, Ottonello S: **Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*.** *J Mol Biol* 1997, **268**:322-330.
28. Hani J, Feldmann H: **tRNA genes and retroelements in the yeast genome.** *Nucleic Acids Res* 1998, **26**:689-696.
29. Lowe TM, Eddy SR: **A computational screen for methylation guide • snoRNAs in yeast.** *Science* 1999, **283**:1168-1171.
This is an excellent example of the 'new' computational biology, where the computer is used to predict a function encoded in the DNA, and the prediction is then tested by direct experiment. The authors identified nearly all of the snoRNAs in yeast, which guide the methylation of rRNA.
30. Olivas WM, Muhlrud D, Parker R: **Analysis of the yeast genome: identification of new non-coding and small ORF-containing RNAs.** *Nucleic Acids Res* 1997, **25**:4619-4625.
31. Schuldiner O, Yanover C, Benvenisty N: **Computer analysis of the entire budding yeast genome for putative targets of the GCN4 transcription factor.** *Curr Genet* 1998, **33**:16-20.
32. Zhong H, McCord R, Vershon AK: **Identification of target sites of the alpha2-Mcm1 repressor complex in the yeast genome.** *Genome Res* 1999, **9**:1040-1047.
33. Hughes JD, Estep PW, Tavazoie S, Church GM: **Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*.** *J Mol Biol* 2000, **296**:1205-1214.
34. Jensen LJ, Knudsen S: **Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation.** *Bioinformatics* 2000, **16**:326-333.
35. van Helden J, Andre B, Collado-Vides J: **A web site for the computational analysis of yeast regulatory sequences.** *Yeast* 2000, **16**:177-187.
36. Hu YJ, Sandmeyer S, McLaughlin C, Kibler D: **Combinatorial motif analysis and hypothesis generation on a genomic scale.** *Bioinformatics* 2000, **16**:222-232.
37. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: **Systematic determination of genetic network architecture.** *Nat Genet* 1999, **22**:281-285.
38. Brazma A, Jonassen I, Vilo J, Ukkonen E: **Predicting gene regulatory elements *in silico* on a genomic scale.** *Genome Res* 1998, **8**:1202-1215.
39. Roth FP, Hughes JD, Estep PW, Church GM: **Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation.** *Nat Biotechnol* 1998, **16**:939-945.
40. van Helden J, Andre B, Collado-Vides J: **Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies.** *J Mol Biol* 1998, **281**:827-842.
41. Anderson JSJ, Parker R: **Computational identification of cis-acting elements affecting post-transcriptional control of gene expression in *Saccharomyces cerevisiae*.** *Nucleic Acids Res* 2000, **28**:1604-1617.
42. van Helden J, del Olmo M, Perez-Ortin JE: **Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals.** *Nucleic Acids Res* 2000, **28**:1000-1010.
43. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95**:14863-14868.
44. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D: **A combined algorithm for genome-wide prediction of protein function.** *Nature* 1999, **402**:83-86.
45. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H *et al.*: **Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis.** *Science* 1999, **285**:901-906.
46. Niedenthal R, Riles L, Guldener U, Klein S, Johnston M, Hegemann JH: **Systematic analysis of *S. cerevisiae* chromosome VIII genes.** *Yeast* 1999, **15**:1775-1796.
47. Rieger KJ, El-Alama M, Stein G, Bradshaw C, Slonimski PP, Maundrell K: **Chemotyping of yeast mutants using robotics.** *Yeast* 1999, **15**:973-986.
48. Entian KD, Schuster T, Hegemann JH, Becher D, Feldmann H, Guldener U, Gotz R, Hansen M, Hollenberg CP, Jansen G *et al.*: **Functional analysis of 150 deletion mutants in *Saccharomyces cerevisiae* by a systematic approach.** *Mol Gen Genet* 1999, **262**:683-702.
49. Shoemaker DD, Lashkari DA, Morris D, Mittmann M, Davis RW: **Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy.** *Nat Genet* 1996, **14**:450-456.
50. Smith V, Chou KN, Lashkari D, Botstein D, Brown PO: **Functional analysis of the genes of yeast chromosome V by genetic footprinting.** *Science* 1996, **274**:2069-2074.
51. Giaever G, Shoemaker DD, Jones TW, Liang H, Winzeler EA, • Astromoff A, Davis RW: **Genomic profiling of drug sensitivities via induced haploinsufficiency.** *Nat Genet* 1999, **21**:278-283.
The authors employ an approach to find genes involved in particular functions that is underutilized in yeast: searching for haploinsufficient genes. The authors identify proteins that are targets of some select drugs by looking throughout the genome for genes that increase sensitivity to the drug when their copy number in the cell cut in half (i.e. in a heterozygous diploid mutant). The key to the approach is the use of 'molecular bar codes' that mark individual mutants, which enables first, screening a large number of mutants (potentially all 6000 yeast gene deletion mutants) in one experiment, and second, detection of small effects of the drug on growth of the mutant.
52. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, • Armour CD, Bennett HA, Coffey E, Dai H, He YD *et al.*: **Functional discovery via a compendium of expression profiles.** *Cell* 2000, **102**:109-126.
The authors measure expression of all ~6000 genes in the yeast genome in each of about 300 different mutants. The set of genes whose expression changes in the mutants serves as a phenotype (a 'fingerprint') of that mutant. The fingerprints provide a database of such phenotypes (a 'compendium') that often allows prediction of protein function. For example, the authors identified a new gene (*ERG28/YER044*) involved in ergosterol biosynthesis because the mutant exhibited a genome expression fingerprint similar to the fingerprint of a mutant defective in a known *ERG* gene. A database of such genome expression fingerprints of all 6000 yeast mutants would likely suggest the function of a substantial number of the yeast genes of unknown function.
53. Rine J: **Gene overexpression in studies of *Saccharomyces cerevisiae*.** *Methods Enzymol* 1991, **194**:239-251.
54. Ramer SW, Elledge SJ, Davis RW: **Dominant genetics using a yeast genomic library under the control of a strong inducible promoter.** *Proc Natl Acad Sci USA* 1992, **89**:11589-11593.
55. Derisi JL, Iyer VR, Brown PO: **Exploring the metabolic and genetic control of gene expression on a genomic scale.** *Science* 1997, **278**:680-686.
56. Wodicka L, Dong H, Mittmann M, Ho MH, Lockhart DJ: **Genome-wide expression monitoring in *Saccharomyces cerevisiae*.** *Nat Biotechnol* 1997, **15**:1359-1367.
57. Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA: **Dissecting the regulatory circuitry of a eukaryotic genome.** *Cell* 1998, **95**:717-728.
58. Wyrick JJ, Holstege FC, Jennings EG, Causton HC, Shore D, Grunstein M, Lander ES, Young RA: **Chromosomal landscape of nucleosome-dependent gene expression and silencing in yeast.** *Nature* 1999, **402**:418-421.

59. Lyons TJ, Gasch AP, Gaitner LA, Botstein D, Brown PO, Eide DJ: **Genome-wide characterization of the Zap1p zinc-responsive regulon in yeast.** *Proc Natl Acad Sci USA* 2000, **97**:7957-7962.
60. DeRisi J, van den Hazel B, Marc P, Balzi E, Brown P, Jacq C, Goffeau A: **Genome microarray analysis of transcriptional activation in multidrug resistance yeast mutants.** *FEBS Lett* 2000, **470**:156-160.
61. Sudarsanam P, Iyer VR, Brown PO, Winston F: **Whole-genome expression analysis of snf/swi mutants of *Saccharomyces cerevisiae*.** *Proc Natl Acad Sci USA* 2000, **97**:3364-3369.
62. Lutfiyya LL, Iyer VR, DeRisi J, DeVit MJ, Brown PO, Johnston M: **Characterization of three related glucose repressors and genes they regulate in *Saccharomyces cerevisiae*.** *Genetics* 1998, **150**:1377-1391.
63. Posas F, Chambers JR, Heyman JA, Hoefler JP, de Nadal E, Arino J: **The transcriptional response of yeast to saline stress.** *J Biol Chem* 2000, **275**:17249-17255.
64. Roberts CJ, Nelson B, Marton MJ, Stoughton R, Meyer MR, Bennett HA, He YD, Dai H, Walker WL, Hughes TR *et al.*: **Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles.** *Science* 2000, **287**:873-880.
65. Jung US, Levin DE: **Genome-wide analysis of gene expression regulated by the yeast cell wall integrity signalling pathway.** *Mol Microbiol* 1999, **34**:1049-1057.
66. Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO, Herskowitz I: **The transcriptional program of sporulation in budding yeast.** *Science* 1998, **282**:699-705.
67. Gygi SP, Rochon Y, Franza BR, Aebersold R: **Correlation between protein and mRNA abundance in yeast.** *Mol Cell Biol* 1999, **19**:1720-1730.
68. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P *et al.*: **A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*.** *Nature* 2000, **403**:623-627.
69. Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, Yamamoto K, Kuhara S, Sakaki Y: **Toward a protein-protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins.** *Proc Natl Acad Sci USA* 2000, **97**:1143-1147.
70. Fromont-Racine M, Mayes AE, Brunet-Simon A, Rain JC, Colley A, Dix I, Decourty L, Joly N, Ricard F, Beggs JD *et al.*: **Genome-wide protein interaction screens reveal functional networks involving Sm-like proteins.** *Yeast* 2000, **17**:95-110.
71. Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, Seraphin B: **A generic protein purification method for protein complex characterization and proteome exploration.** *Nat Biotechnol* 1999, **17**:1030-1032.

This paper describes an apparently robust method for purifying protein complexes, the components of which can then be identified by mass spectrometry. The authors developed a double epitope tag that allows for purification of protein complexes to near homogeneity. They used the method to identify some protein complexes involved in mRNA splicing.

72. Martzen MR, McCraith SM, Spinelli SL, Torres FM, Fields S, Grayhack EJ, Phizicky EM: **A biochemical genomics approach for identifying genes by the activity of their products.** *Science* 1999, **286**:1153-1155.

The authors have made biochemical analysis of proteins accessible to even the most hard-core yeast geneticist (e.g. me!) They fused each of the 6000 yeast proteins to a 'purification hook' (in this case glutathione-S-transferase [GST]), thereby enabling their rapid and easy purification (by binding to glutathione conjugated to agarose beads). One can now, in principle, test all 6000 yeast proteins for an activity of interest in a very short time. The authors identified several enzymes involved in tRNA processing in this way.