

anti-leukaemia treatments have historically sought to eliminate proliferating cells. But, like normal HSCs, L-HSCs rarely proliferate and so are resistant to such therapies. The identification of L-HSC-specific genetic targets, especially those involved in the enhanced self-renewal seen in L-HSCs, could lead to new treatments that stop the disease at its source. ■

John E. Dick is in the Division of Cell and Molecular Biology, University Health Network, and the Department of Molecular Genetics and Microbiology, University of Toronto, 620 University Avenue, Toronto, Ontario M5G 2C1, Canada. e-mail: jdick@uhnres.utoronto.ca

1. Kondo, M. *et al.* *Annu. Rev. Immunol.* **21**, 759–806 (2003).
2. Devine, S. M., Lazarus, H. M. & Emerson, S. G. *Bone Marrow*

3. Lessard, J. & Sauvageau, G. *Nature* **423**, 255–260 (2003).
4. Park, I. *et al.* *Nature* **423**, 302–305 (2003).
5. Till, J. E. & McCulloch, E. A. *Radiat. Res.* **14**, 213–222 (1961).
6. Siminovitch, L., McCulloch, E. A. & Till, J. E. *J. Cell. Comp. Physiol.* **62**, 327–336 (1963).
7. Till, J. E., McCulloch, E. A. & Siminovitch, L. *Proc. Natl Acad. Sci. USA* **51**, 29–36 (1964).
8. Reya, T. *et al.* *Nature* advance online publication, 27 April 2003 (doi:10.1038/nature01593).
9. Willert, K. *et al.* *Nature* advance online publication, 27 April 2003 (doi:10.1038/nature01611).
10. Sauvageau, G. *et al.* *Genes Dev.* **9**, 1753–1765 (1995).
11. van Lohuizen, M. *et al.* *Cell* **65**, 737–752 (1991).
12. van der Lugt, N. *et al.* *Genes Dev.* **8**, 757–769 (1994).
13. Dick, J. E. *Proc. Natl Acad. Sci. USA* **100**, 3547–3549 (2003).
14. Lapidot, T. *et al.* *Nature* **367**, 645–648 (1994).
15. Wang, J. C. *et al.* in *Hematopoiesis: A Developmental Approach* (ed. Zon, L. I.) 99–118 (Oxford Univ. Press, New York, 2001).
16. Bonnet, D. & Dick, J. E. *Nature Med.* **3**, 730–737 (1997).
17. Kroon, E. *et al.* *EMBO J.* **17**, 3714–3725 (1998).
18. Iscove, N. N. & Nawa, K. *Curr. Biol.* **7**, 805–808 (1997).

## Genomics

# Yeast rises again

Steven L. Salzberg

What's in a genome? The short answer is that you can't really say in detail for any one species until you have the genome sequences of a variety of other species — some closely related, others less so — to compare it with.

While the human and mouse genomes have lately dominated public discussions of genome science, yeast researchers have quietly continued to forge ahead with analysing *Saccharomyces cerevisiae* and related yeast species. Yeast has long been a favourite of biologists, who use it as a model for investigating the biology of higher organisms. Although yeasts are unicellular, like bacteria, they have a cell nucleus, putting them in the biological kingdom of eukaryotes (Eukaryota), a group that includes humans. Yeast was the first eukaryote to have its genome completely sequenced, in 1996<sup>1</sup>. And yeast has long been a part of human commerce and culture, being used in brewing and baking since ancient times<sup>2</sup>. *Saccharomyces cerevisiae* itself was first isolated in beer in 1837.

On page 241 of this issue, Kellis and colleagues<sup>3</sup> describe the results of sequencing three more yeast species — *Saccharomyces paradoxus*, *S. bayanus* and *S. mikatae* — and comparing them with *S. cerevisiae*. Two of these species, *S. bayanus* and *S. paradoxus*, are used in winemaking, but the three new species were chosen mainly because they are evolutionarily very close to *S. cerevisiae*, and the implications of the study go well beyond brewing and viticulture. The analyses will produce a substantial revision in our knowledge of the yeast genome, and provide strategic directions for how we might select other sequencing targets to advance understanding of the human genome.

The original *S. cerevisiae* project was a model of international collaboration among small laboratories: 55% of the genome was

sequenced by a network of European labs, with the rest of it being generated at five large (at the time) centres. More than 600 scientists participated in the project, sequencing a cosmid (a 40-kilobase fragment) at a time and finishing the sequences to a high level of accuracy with minimal automation<sup>1</sup>. In the seven years since then, genome sequencing has moved increasingly into large-scale factory operations: the three new genomes were produced at a single centre using the whole-genome shotgun strategy, which is much more rapid than the earlier piecemeal sequencing approaches.

The capacity of today's largest centres allows them to sequence a yeast genome — 12 million base pairs or so — in just a few days of full-scale production. Developments

in genome assemblers<sup>4,5</sup> make it possible to assemble the fragments produced by the whole-genome shotgun approach into relatively large pieces with few mistakes. This assembly step works even better when a reference genome already exists: the assembled fragments can be quickly mapped onto the reference, assuming that relatively few rearrangements have occurred since the species diverged. As shown in Fig. 2 of Kellis and colleagues' paper (page 244), all 16 chromosomes from each of the three newly sequenced genomes map beautifully onto *S. cerevisiae*.

The main reason for sequencing the three further yeast species is to better understand *S. cerevisiae*. A few years ago, Cliften and colleagues<sup>6</sup> and the French consortium Génolevures<sup>2</sup> independently conducted studies that demonstrated the power of sequencing close relatives of *S. cerevisiae*. Both groups recommended full-scale sequencing of additional yeast species, which they argued would result in the discovery of noncoding RNA genes, genes encoding small proteins, and gene-regulatory sites. In a spot-on prediction, Cliften *et al.* estimated that at least 40 small proteins (shorter than 100 amino acids) would be identified by comparing *S. cerevisiae* with *S. bayanus*, *S. mikatae* and a third species. Kellis and colleagues found 43.

Comparative genome sequencing also helps to resolve the surprisingly difficult question of just how many genes there are. Genes — those parts of a DNA sequence that encode proteins or functional RNA molecules — are the basic components that do all the work of a cell, from reproduction to metabolism to inter-cell communication. Nearly seven years after the yeast genome was first published, the gene count stands at approximately 6,128 (refs 7–9), only 150 or so fewer than the initial estimate. But many of these genes are still questionable: the Génolevures programme<sup>10</sup>, after analysing sequences sampled from 13 different yeast

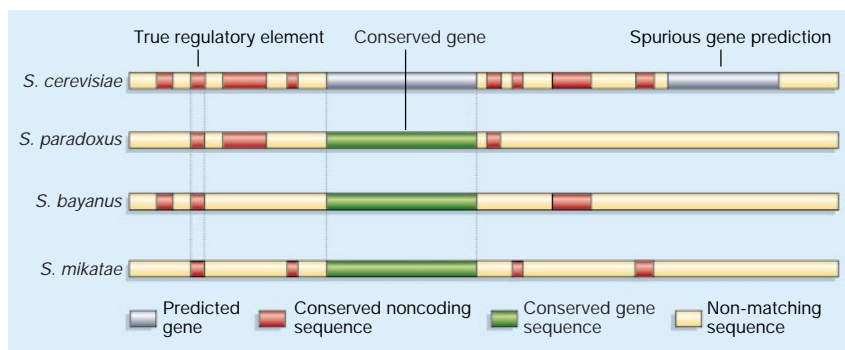


Figure 1 Comparative genomics. Comparing the DNA sequences from several species makes it possible to find regulatory regions — short sequences that turn genes on and off — and eliminate spurious gene predictions. Red boxes highlight areas of sequence similarity between at least two species. Functional sequences — genes and regulatory elements — tend to be conserved across all species. The figure shows how one true regulatory element and one correctly identified gene might emerge from a comparison of four yeast species.

species (including *S. bayanus*), and failing to find matches for 742 genes, estimated that the gene count should be reduced to 5,651. The crux of their argument was the assumption that if a gene is functional, it should be conserved among closely related species (see Fig. 1). Because the Génolevures project covered only 20–40% of each genome, the possibility remained that those ‘missing’ genes might be found in the unsequenced regions.

Kellis and colleagues<sup>3</sup> have now closed the door on that possibility: their sequence data cover 98% of two species and 93% of the third. Based on sequence alignments among the species, they largely confirm the results of the earlier study and argue that 503 genes should be deleted from the yeast catalogue, leaving 5,726 genes, of which 43 are newly discovered in their study.

Regulatory sequences, which sit outside genes and turn them on and off, are the key to understanding how a genome fits together. Whether we are comparing human and mouse, or yeast and yeast, we still have to answer the puzzling question of how seemingly huge differences in physical, biochemical or behavioural characteristics can result from sometimes tiny differences in the protein sequences. Regulatory sites occur virtually anywhere in the vast areas between protein-coding regions, and they can be identified because — unlike non-functional regions — they are conserved. The closer two species are, the more regulatory sites they are likely to share. Unfortunately, if the species are close enough, many pieces of non-functional DNA will be conserved merely by chance. A nice solution to this problem is to sequence more than two related species, dramatically increasing the signal-to-noise ratio (Fig. 1). The idea is that functional sequences should be conserved across multiple species, whereas chance conservation will only appear in pairwise comparisons. Using this principle, Kellis and colleagues have identified 42 novel sequence motifs that appear likely to have biological functions in yeast.

The most dynamic parts of the yeast genome are the chromosome ends, called telomeres. Kellis *et al.* aptly describe the rapid change and exchange going on in these regions as “genomic churning”. The telomeres contain many genes not found elsewhere in the genome, and it appears that they form a crucible in which genomic change occurs: as the telomeres swap back and forth between chromosomes, they can carry pieces of genes along with them, which may combine with others to create new genes. Similarly rapid changes are evident in the telomeres of the malaria parasite *Plasmodium falciparum*<sup>11</sup>. Given the important events associated with these regions, sequencing of telomeres of the human genome — which, so far, have been neglected — should become a priority.

If 8% of the estimated protein complement

of yeast was wrong, how much of the human counterpart might be eliminated by a similar study of mammalian species? How many genes and regulatory regions would be discovered? The human gene count has already dropped from more than 31,000 to just under 25,000 (refs 12, 13) in the two years since the initial genome publication. To understand our own genome better, we should sequence the genomes of several other mammals besides mouse and rat (both are almost done), choosing species that vary in their evolutionary distance from humans. At the chromosome level, only primates have as much structural similarity to human as these four yeast species share, and even those have different chromosome numbers: chimpanzee (*Pan troglodytes*) has 24, gibbon (*Hylobates concolor*) has 25, and macaque (*Macaca fuscata*) has 21. At the sequence level, we might gain more information from studying more distant mammals, such as cat (*Felis catus*), pig (*Sus scrofa*) and dolphin (*Tursiops truncatus*).

This new study of yeast genomes<sup>3</sup> makes

#### Molecular biology

## Disruptive influence

Marco Foiani

Recombination is a vital cellular process implicated in DNA metabolism — but it must be tightly controlled. The discovery of a protein that disrupts recombination intermediates sheds light on the control mechanisms.

On pages 305 and 309 of this issue, Krejci and co-workers<sup>1</sup> and Veaute and colleagues<sup>2</sup> describe a biochemical mechanism that controls the genome ‘shuffling’ occurring in dividing cells and in DNA repair. Their findings have implications for how genome stability is maintained, and hence for the development of cancer.

Genome shuffling is referred to as ‘recombination’, and is a cellular process by which extensive tracts of DNA are moved from one part of the genome to another. There are several recombination pathways<sup>3</sup>, some not yet well characterized, which are routinely used by normal cells to repair damaged chromosomes, to assist in DNA synthesis, and even to regulate gene expression. Recombination also occurs during the production of eggs and sperm, in which its function is to mix the genetic information such that each egg and each sperm is genetically different.

Despite its importance, however, recombination can sometimes be harmful: it can generate damaging genomic rearrangements, as well as intermediate structures that cannot be processed normally. Cells need to coordinate recombination with other responses to DNA damage, with progression through the cell-division cycle, and with

it clear that comparative genome sequencing has tremendous analytical power: it offers the prospect of enhancing our knowledge of thousands of genes at once as well as providing fresh clues about the function of the vast amount of genomic DNA that does not encode genes. As it has done before, lowly yeast shows us a path towards a better understanding of our own biology.

Steven L. Salzberg is at *The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA.*  
e-mail: [salzberg@tigr.org](mailto:salzberg@tigr.org)

- Goffeau, A. *et al. Science* **274**, 546, 563–567 (1996).
- Feldmann, H. *FEBS Lett.* **487**, 1–2 (2000).
- Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. *Nature* **423**, 241–254 (2003).
- Myers, E. W. *et al. Science* **287**, 2196–2204 (2000).
- Batzoglou, S. *et al. Genome Res.* **12**, 177–189 (2002).
- Clifften, P. F. *et al. Genome Res.* **11**, 1175–1186 (2001).
- Snyder, M. & Gerstein, M. *Science* **300**, 258–260 (2003).
- <http://genome-www.stanford.edu/Saccharomyces>
- <http://mips.gsf.de/proj/yeast/CYGD/db>
- Malpertuy, A. *et al. FEBS Lett.* **487**, 113–121 (2000).
- Gardner, M. J. *et al. Nature* **419**, 498–511 (2002).
- International Human Genome Sequencing Consortium *Nature* **409**, 860–921 (2001).
- [www.ensembl.org](http://www.ensembl.org)