Current Biology, Vol. 12, R294–R296, April 16, 2002, ©2002 Elsevier Science Ltd. All rights reserved. PII S0960-9822(02)00805-9

Genome Sequencing: And Then There Were Six

Dispatch

Stuart A. MacNeill

The genome of the fission yeast *Schizosaccharomyces pombe* has been sequenced, bringing the number of sequenced eukaryotic genomes to six. Analysis of the sequence predicts only 4824 protein coding genes, the smallest number yet recorded for a free-living eukaryote.

The fission yeast *Schizosaccharomyces pombe* has proved to be an excellent model system for many eukaryotic cellular processes, in particular cell-cycle control, chromosome structure, cytoskeletal organisation and function, and mitosis. Now the genome of this free-living unicellular ascomycete has been sequenced and annotated [1], providing fascinating insights into the biology of this model organism, together with a wealth of new data for bioinformatic analysis. *S. pombe* is the sixth free-living eukaryote to have its genome fully sequenced, and the second yeast, after the budding yeast *Saccharomyces cerevisiae* in 1996 [2,3]. In evolutionary terms, the two yeasts are only distantly related, with recent estimates [4] suggesting that they diverged from one another ~1,000–1,200 million years ago, and from metazoa ~1,600 million years ago, significantly earlier than was previously thought [5]. The completion of the fission yeast genome sequence marks a significant step forward in the life of this model eukaryote and heralds the start of an exciting period for fission yeast research.

The fission yeast genome totals ~13.8 megabases, distributed between three chromosomes of 5.7, 4.6 and 3.5 megabases [1]. To sequence the genome, pre-existing genome-wide cosmid maps [6,7] were refined and corrected to generate a minimal tile path across the genome. Cosmid sequencing was then carried out by an international consortium headed by Paul Nurse (Cancer Research UK) and Bart Barrell (Wellcome Trust Sanger Institute). Gaps in the cosmid-derived sequence were filled using long-range polymerase chain reaction (PCR), bacterial artificial chromosome (BAC) clones and genomic DNA library plasmids. In total, 452 cosmids, 22 plasmids, 15 BACs and 13 PCR products were sequenced to generate the genome sequence with an average depth of coverage of approximately eight-fold [1]. Satisfyingly, the published sequence contains few gaps: those that remain are located in the highly repetitive centromeric and telomeric regions, and in the ribosomal (r)DNA region of chromosome III that comprises 100–120 tandem copies of the 5.8S, 18S and 25S rRNA genes. All the unique regions of the genome have been sequenced, totalling 12.5 megabases.

The availability of the whole genome sequence provides significant insights into how chromosome organisation and gene structure in *S. pombe* compares to that in *S. cerevisiae*. The fission yeast genome is less gene dense than that of its distant relative and intragenic regions are typically longer, perhaps reflecting more complex promoter structures. In addition, introns are commonplace in fission yeast genes: 43% of fission yeast genes contain introns, as compared to only 5% in budding yeast. As is the case in *S. cerevisiae*, introns tend to be located towards the 5′ end of open reading frames, consistent with models for intron loss resulting from gene conversion by homologous recombination using reverse transcribed mRNAs [8]. As Nurse and co-workers point out [1], with so many spliced genes, the potential for alternative splicing to be used to increase the range of protein species available to the cell is also greater.

The new data also provide a detailed view of the structure of the *S. pombe* centromeres [1]. As shown by the earlier work by the Yanagida and Clarke labs, fission yeast centromeres are considerably larger and more complex than their budding yeast counterparts, extending over 35, 65 and 110 kilobases on chromosomes I, II and III respectively — the reason for the apparent inverse relationship between the length of the centromere and that of the corresponding chromosome is not known — and are thought to be an excellent model for higher eukaryotic centromeres. Each centromere contains a central core region surrounded by numerous repeated sequence elements in an extended region that is devoid of protein coding potential, although several tRNA genes are found there. The precise function of these centromere-embedded tRNA genes is unclear, but they may act as boundary elements, delineating the edges of functionally important chromatin domains [9].

Analysis of the fission yeast genome identifies at least 4,824 protein coding genes, including eleven encoded by the mitochondrial genome and a possible 33 pseudogenes [1], and perhaps as many as 4,940, although this higher figure includes 116 genes that Nurse and co-workers [1] describe (for various reasons) as questionable. For comparison, the *Arabidopsis*, *Caenorhabditis elegans* and *Drosophila* genomes contain 27,000, 19,100 and 13,600 genes, in genomes of 120, 100 and 180 megabases, respectively [10–12]; the vast, 3.2 gigabase human genome probably carries 30,000–40,000 genes [13,14] (see Figure 1). Interestingly, the number of fission yeast genes is significantly less than the 5,800 or so predicted for *S. cerevisiae* [2,3].

Comparisons of the proteomes of the two yeasts and *C. elegans* (Figure 2) are revealing [1]. Over two-thirds of *S. pombe* proteins — 3,281, 67% — have

Wellcome Trust Centre for Cell Biology, University of Edinburgh, King's Buildings, Edinburgh EH9 3JR, UK.
E-mail: s.a.macneill@ed.ac.uk

homologues in *S. cerevisiae* and *C. elegans*, while only 681 (14%) are unique to *S. pombe*. A further 769 (16%) are found in *S. pombe* and *S. cerevisiae*, but not *C. elegans*, while 145 (< 3%) are found in *S. pombe* and *C. elegans* but not *S. cerevisiae*. A similar analysis of the *S. cerevisiae* proteome (Figure 2, lower part) reveals an almost identical pattern, except that there are more *S. cerevisiae* proteins with homologues in *S. pombe* (918) than there are *S. pombe* proteins with homologues in *S. cerevisiae* (769) [1].

The explanation for this lies in the number of gene families in the two yeasts. Nurse and co-workers [1] report that ~93% of *S. pombe* genes — 4,515 from a data set of 4,876 — can be considered as being unique genes, that is, they have no other sequence relatives within the fission yeast genome. The remaining 7% (361) are distributed among protein cluster groups, with two or more family members per group. In *S. cerevisiae*, however, almost twice as many genes (716, or 12% of the total) fall into the latter category. Put simply, there is greater redundancy in the budding yeast genome, consistent with the occurrence of large-scale genome duplication events during the evolution of this organism [2,3].

With fewer than 5,000 protein-coding genes, *S. pombe* has fewer genes than a number of eubacterial organisms whose genomes have been sequenced (see Figure 1). The 6.3 megabase genome of the opportunistic pathogen *Pseudomonas aeruginosa*, for example, encodes 5,570 proteins [15], while the 8.7 megabase genome of the industrial microorganism *Streptomyces avermitilis* encodes at least 7,600 [16]. Clearly, there is no direct correlation between gene number and being a prokarote or a eukaryote. What then defines the eukaryotic cell? Nurse and co-workers [1] present an initial analysis of this problem by comparing the proteins encoded by the six sequenced eukaryotic genomes with those of thirty-seven eubacterial and eight archaeal species. By identifying proteins that are well conserved across different eukaryotic species (with protein sequence similarity greater than or equal to 50%) but which are poorly, or not at all, conserved in prokaryotes (protein sequence similarity of less than or equal to 20%), they identify some 62 proteins that distinguish eukaryotes from prokaryotes, and classify these into eight functional groups.

The identity of these groups makes interesting reading [1]. Not surprisingly, one group contains ribosomal proteins — recall that eukaryotic ribosomes are larger than prokaryotic ones — while another contains histone H3 and H4 proteins involved in packing DNA in nucleosomes, another characteristic feature of eukaryotic cells. Additional groups include cytoskeletal proteins —actin, tubulin and so on — as well as proteins involved in subcellular compartmentalisation, cell-cycle control (including the Cdc2 protein kinase), protein phosphorylation and dephosphorylation, regulated proteolysis and splicing. In short, many of the proteins identified by this analysis are involved in processes that we consider to be characteristic of eukaryotic life.

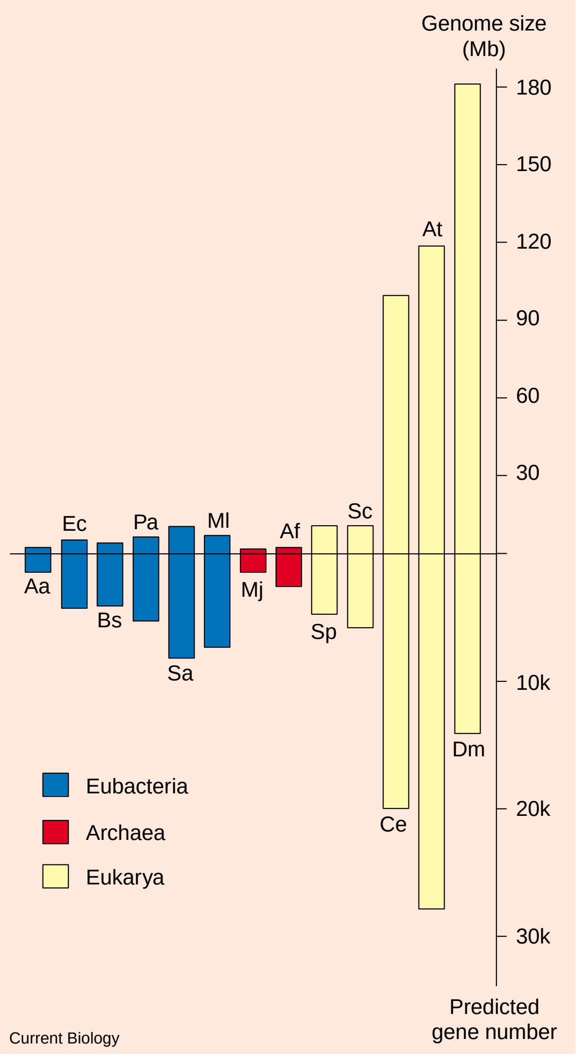

Figure 1. Genome size and gene number of sequenced genomes.

Upper part: sizes of the sequenced genomes of six eubacterial (blue), two archaeal (red) and five eukaryotic (yellow) species. Lower part: estimated number of protein-coding genes in each species. Abbreviations, genome sizes and estimated gene numbers: Aa (*Aquifex aeolicus*, 1.55 Mb, 1,512), Ec (*E. coli* K12, 4.67 Mb, 4,288), Bs (*Bacillus subtilis*, 4.21 Mb, 4,100), Pa (*Pseudomonas aeruginosa*, 6.3 Mb, 5,570), Sa (*Streptomyces avermitilis*, 8.7 Mb, >7,600), Ml (*Mesorhizobium loti*, 7.0 Mb, 6,752), Mj (*Methanoccus jannaschii*, 1.67 Mb, 1,738), Af (*Archaeoglobus fulgidus*, 2.18 Mb, 2,436), *S. pombe* (12.5 Mb, 4,940), *S. cerevisiae* (12.5 Mb, 5,777), Ce (*C. elegans*, 97 Mb, 19,099), At (*A.thaliana*, 120 Mb, 27,000) and Dm (*Drosophila melanogaster*, 180 Mb, 13,600). Note that, in the case of *Mesorhizobium loti*, the genome size and gene number relate to the chromosome only and does not include the megaplasmids, and that in the case of *Drosophila*, the so-called complete genome sequence does not include ~60 Mb heterochromatic sequences [11]. The human genome is thought to be ~3.2 Gb in length but with only 30,000–40,000 protein coding genes [13,14].

A similar analysis was performed to identify genes important for multicellular, rather than unicellular life [1]. In this case, the proteins encoded by the

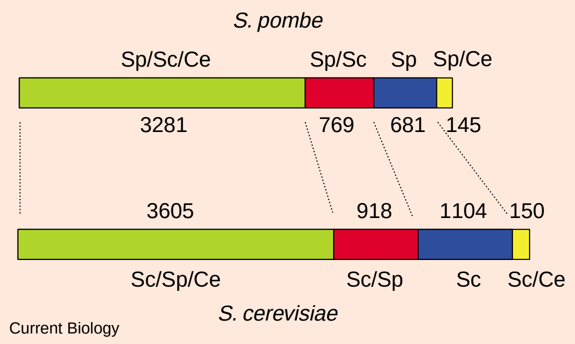

Figure 2. Evolutionary conservation of yeast proteins.

Top: 3,281 *S. pombe* (Sp) proteins (represented as a green box) have homologues in *S. cerevisiae* (Sc) and *C. elegans* (Ce), a further 769 have homologues in *S. cerevisiae* alone (red box), 681 are unique to *S. pombe* (blue box), and 145 have homologues in *C. elegans* but not *S. cerevisiae* (yellow box). Note that the dataset used for this analysis comprised 4,876 *S. pombe* proteins only [1]. Bottom: a similar analysis for *S. cerevisiae*. See text for discussion.

unicellular organisms — eubacteria, archaea and yeast — were grouped in a single data set and compared to proteins encoded by the genomes of the multicellular eukaryotes. The result of this analysis was striking: only a very small number of proteins were identified — one to three, depending on the protein sequence similarity threshold applied. It is difficult to escape the conclusion that the transition from unicellular to multicellular life did not involve the evolution of many new proteins, but involved instead the adaptation of pre-existing protein types. As an example of this type of process, Nurse and colleagues [1] highlight the way in which the basic cell–cell signalling machinery developed by the yeasts for seeking out mating partners has been adapted to facilitate intercellular signalling in multicellular organisms.

The completion of the *S. pombe* genome sequence is a landmark event in the history of fission yeast research. In addition to reinforcing the reputation of this single-celled eukaryote as an excellent model system for understanding higher eukaryotic cellular processes, such as cell-cycle control, chromosome structure and mitosis, the availability of genome sequence will doubtless enhance the accessibility of the organism for hitherto neglected areas of eukaryotic cell biology. The future of the fission yeast looks very bright indeed.

References

1. Wood, V., Gwilliam, R., Rajandream, M.-A., Lyne, M., Stewart, A., Sgouros, J., Peat, N., Hayles, J., Baker, S., Basham, D. *et al.* (2002). The genome sequence of *Schizosaccharomyces pombe*. Nature *415*, 871–880.
2. Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M. *et al.* (1996). Life with 6000 genes. Science *546*, 563–567.
3. Goffeau, A., Aert, R., Agostini-Carbone, M.L., Ahmed, A., Aigle, M., Alberghina, L., Albermann, K., Albers, M., Aldea, M., Alexandraki, D. *et al.* (1997). The Yeast Genome Directory. Nature *387* (suppl.), 1–105.
4. Heckman, D.S., Geiser, D.M., Eidell, B.R., Stauffer, R.L., Kardos, N.L. and Hedges, S.B. (2001). Molecular evidence for the early colonization of land by fungi and plants. Science *293*, 1129–1133.
5. Sipiczki, M. (2000). Where does fission yeast sit on the tree of life? Genome Biol. *1*, 1011.
6. Hoheisel, J.D., Maier, E., Mott, R., McCarthy, L., Grigoriev, A.V., Schalkwyk, L.C., Nizetic, D., Francis, F. and Lehrach, H. (1993). High resolution cosmid and P1 maps spanning the 14 Mb genome of the fission yeast *S. pombe*. Cell *73*, 109–120.
7. Mizukami, T., Chang, W.I., Garkavtsev, I., Kaplan, N., Lombardi, D., Matsumoto, T., Niwa, O., Kounosu, A., Yanagida, M. and Marr, T.G. (1993). A 13 kb resolution cosmid map of the 14 Mb fission yeast genome by non-random sequence-tagged site mapping. Cell *73*, 121–123.
8. Fink, G.R. (1987). Pseudogenes in yeast? Cell *49*, 5–6.
9. Partridge, J.F., Borgstrom, B. and Allshire R.C. (2000). Distinct protein interaction domains and protein spreading in a complex centromere. Genes Dev. *14*, 783–791.
10. The *C. elegans* Sequencing Consortium. (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science *282*, 2012–2018.
11. Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F. *et al.* (2000). The genome sequence of *Drosophila melanogaster*. Science *287*, 2185–2195.
12. The Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature *408*, 796–813.
13. International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.
14. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. *et al.* (2001). The sequence of the human genome. Science *291*, 1304–1351.
15. Stover, C.K., Pham, X.Q., Erwin, A.L., Mizoguchi, S.D., Warrener, P., Hickey, M.J., Brinkman, F.S.L., Hufnagle, W.O., Kowalik, D.J., Lagrou, M., *et al*. (2000). Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. Nature *406*, 959–964.
16. Omura, S., Ikeda, H., Ishikawa, J., Hanamoto, A., Takahashi, C., Shinose, M., Takahashi, Y., Horikawa, H., Nakazawa, H., Osonoe, T., *et al.* (2001). Genome sequence of an industrial microorganism *Streptomyces avermitilis*: deducing the ability of producing secondary metabolites. Proc. Natl. Acad. Sci. U.S.A. *98*, 12215–12220.