The genome sequence of *Schizosaccharomyces pombe*

V. Wood[1], R. Gwilliam[1], M.-A. Rajandream[1], M. Lyne[1], R. Lyne[1], A. Stewart[2], J. Sgouros[2], N. Peat[3], J. Hayles[3], S. Baker[1], D. Basham[1], S. Bowman[1], K. Brooks[1], D. Brown[1], S. Brown[1], T. Chillingworth[1], C. Churcher[1], M. Collins[1], R. Connor[1], A. Cronin[1], P. Davis[1], T. Feltwell[1], A. Fraser[1], S. Gentles[1], A. Goble[1], N. Hamlin[1], D. Harris[1], J. Hidalgo[1], G. Hodgson[1], S. Holroyd[1], T. Hornsby[1], S. Howarth[1], E. J. Huckle[1], S. Hunt[1], K. Jagels[1], K. James[1], L. Jones[1], M. Jones[1], S. Leather[1], S. McDonald[1], J. McLean[1], P. Mooney[1], S. Moule[1], K. Mungall[1], L. Murphy[1], D. Niblett[1], C. Odell[1], K. Oliver[1], S. O'Neil[1], D. Pearson[1], M. A. Quail[1], E. Rabbinowitsch[1], K. Rutherford[1], S. Rutter[1], D. Saunders[1], K. Seeger[1], S. Sharp[1], J. Skelton[1], M. Simmonds[1], R. Squares[1], S. Squares[1], K. Stevens[1], K. Taylor[1], R. G. Taylor[1], A. Tivey[1], S. Walsh[1], T. Warren[1], S. Whitehead[1], J. Woodward[1], G. Volckaert[4], R. Aert[4], J. Robben[4], B. Grymonprez[4], I. Weltjens[4], E. Vanstreels[4], M. Rieger[5], M. Schäfer[5], S. Müller-Auer[5], C. Gabel[5], M. Fuchs[5], C. Fritzc[6], E. Holzer[6], D. Moestl[6], H. Hilbert[6], K. Borzym[7], I. Langer[7], A. Beck[7], H. Lehrach[7], R. Reinhardt[7], T. M. Pohl[8], P. Eger[8], W. Zimmermann[9], H. Wedler[9], R. Wambutt[9], B. Purnelle[10], A. Goffeau[10], E. Cadieu[11], S. Dréano[11], S. Gloux[11], V. Lelaure[11], S. Mottier[11], F. Galibert[11], S. J. Aves[12], Z. Xiang[12], C. Hunt[12], K. Moore[12], S. M. Hurst[12], M. Lucas[13], M. Rochet[13], C. Gaillardin[13], V. A. Tallada[14,15], A. Garzon[14,15], G. Thode[14], R. R. Daga[14,15], L. Cruzado[14], J. Jimenez[14,15], M. Sánchez[16], F. del Rey[16], J. Benito[16], A. Domínguez[16], J. L. Revuelta[16], S. Moreno[16], J. Armstrong[17], S. L. Forsburg[18], L. Cerrutti[1], T. Lowe[19], W. R. McCombie[20], I. Paulsen[21], J. Potashkin[22], G. V. Shpakovski[23], D. Ussery[24], B. G. Barrell[1] & P. Nurse[3]

We have sequenced and annotated the genome of fission yeast (*Schizosaccharomyces pombe*), which contains the smallest number of protein-coding genes yet recorded for a eukaryote: 4,824. The centromeres are between 35 and 110 kilobases (kb) and contain related repeats including a highly conserved 1.8-kb element. Regions upstream of genes are longer than in budding yeast (*Saccharomyces cerevisiae*), possibly reflecting more-extended control regions. Some 43% of the genes contain introns, of which there are 4,730. Fifty genes have significant similarity with human disease genes; half of these are cancer related. We identify highly conserved genes important for eukaryotic cell organization including those required for the cytoskeleton, compartmentation, cell-cycle control, proteolysis, protein phosphorylation and RNA splicing. These genes may have originated with the appearance of eukaryotic life. Few similarly conserved genes that are important for multicellular organization were identified, suggesting that the transition from prokaryotes to eukaryotes required more new genes than did the transition from unicellular to multicellular organization.

We report here the completion of the fully annotated genome sequence of the simple eukaryote *Schizosaccharomyces pombe*, a fission yeast. It becomes the sixth eukaryotic genome to be sequenced, following *Saccharomyces cerevisiae*[1], *Caenorhabditis elegans*[2], *Drosophila melanogaster*[3], *Arabidopsis thaliana*[4] and *Homo sapiens*[5,6]. The entire sequence of the unique regions of the three chromosomes is complete, with gaps in the centromeric regions of about 40 kb, and about 260 kb in the telomeric regions. The completion of this sequence, the availability of sophisticated research methodologies, and the expanding community working on *S. pombe*, will accelerate the use of *S. pombe* for functional and comparative studies of eukaryotic cell processes.

Schizosaccharomyces pombe is a single-celled free living archiascomycete fungus sharing many features with cells of more complicated eukaryotes. From gene sequence comparisons and phylogenetic analyses, it has been suggested that fission yeast diverged from budding yeast around 330–420 million years (Myr) ago, and from Metazoa and plants around 1,000–1,200 Myr ago[7], although a more recent estimate has put these times at 1,144 and 1,600 Myr, respectively[8]. Some gene sequences are as equally diverged between the two yeasts as they are from their human homologues, probably reflecting a more rapid evolution within fungal lineages than in the Metazoa. *S. pombe* was first described in the 1890s and has been extensively studied since the 1950s[9,10], resulting in the characterization of around 1,200 genes (http://www.genedb.org/pombe). The ease with which it can be genetically manipulated is second only to *S. cerevisiae* among eukaryotes and it has served as an excellent model organism for the study of cell-cycle control, mitosis and meiosis[11], DNA repair and recombination[12], and the checkpoint controls important for genome stability[13].

The 13.8-Mb genome of *S. pombe* is distributed between chromosomes I (5.7 Mb), II (4.6 Mb) and III (3.5 Mb)[14], together with a 20-kb mitochondrial genome[15]. Tandem arrays of 100–120 repeats of a 10.4-kb fragment containing the 5.8S, 18S and 25S ribosomal RNA genes account for around 1.1 Mb[16]. The three centromeres are 35, 65 and 110 kb long for chromosomes I, II and III, respectively, totalling 0.2 Mb. This leaves about 12.5 Mb of unique sequence, similar in size to that of *S. cerevisiae*, and substantially smaller than those of the three other sequenced model eukaryotes, *C. elegans* (97 Mb), *Arabidopsis* (125 Mb) and *Drosophila* (137 Mb). All of the

[1]The Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [2]Cancer Research UK London Research Institute, Computational Genome Analysis Laboratory, 44 Lincoln's Inn Fields, London WC2A 3PX, UK. [3]Cancer Research UK London Research Institute, Cell Cycle Laboratory, 44 Lincoln's Inn Fields, London, WC2A 3PX, UK. [4]Katholieke Universiteit Leuven, Faculty of Agricultural and Applied Biological Sciences, Laboratory of Gene Technology, Kardinaal Mercierlaan 92 Blok F, B-3001 Leuven, Belgium. [5]Genotype GmbH, Molecular Biology and Biotech Research, Angelhofweg 39, D-69259 Wilhelmsfeld, Germany. [6]QIAGEN GmbH, Max Volmer Str. 4, D-40724 Hilden, Germany. [7]Max-Planck-Institut für molekulare Genetik, Ihnestrasse 73, D-14195 Berlin, Germany. [8]GATC Biotech AG, Jakob-Stadler-Platz 7, D-78467 Konstanz, Germany. [9]AGOWA GmbH, Glienicker Weg 185, D-12489 Berlin, Germany. [10]Université de Louvain, Unite de Biochimie Physiologique, Place Croix du Sud 2-20, B1348 Louvain-la-Neuve, Belgium. [11]UMR 6061 CNRS Genetique et developpement, Faculté de Médecine, 2 avenue du Professeur Léon Bernard, F-35043 Rennes Cedex, France. [12]University of Exeter, School of Biological Sciences, Washington Singer Laboratories, Perry Road, Exeter EX4 4QG, UK. [13]Génétique Moléculaire et Cellulaire, CNRS URA1925 INRA UMR216, Institut National Agronomique Paris-Grignon, 78850 Thiverval Grignon, France. [14]Departamento de Genetica, Facultad de Ciencias, Universidad de Malaga, Spain. [15]Laboratorio Andaluz de Biologia, Universidad Pablo de Olavide, Sevilla, Spain. [16]Instituto de Microbiología y Bioquímica, Departamento de Microbiología y Genética, CSIC/Universidad de Salamanca, Edificio Departamental, Campus Miguel de Unamuno, 37007 Salamanca, Spain. [17]University of Sussex, Falmer, Brighton BN1 9QG, UK. [18]Molecular & Cell Biology Laboratory, Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, California 92037-1099, USA. [19]Stanford University, Stanford University School of Medicine, Department of Genetics, CCSR Room 2255b, 269 Campus Drive, Stanford, California 94305, USA. [20]Cold Spring Harbor Laboratory, PO Box 100, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA. [21]TIGR, 9712 Medical Center Drive, Rockville, Maryland 20850, USA. [22]The Chicago Medical School, 3333 Green Bay Road, North Chicago, Illinois 60064, USA. [23]Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Ul. Miklukho-Maklaya 16/10, 117997 Moscow, Russia. [24]Center for Biological Sequence Analysis, BioCentrum-DTU, The Technical University of Denmark, Building 208, DK-2800 Kgs. Lyngby, Denmark.

© 2002 Macmillan Magazines Ltd

unique sequence and most of the three centromeres of the Urs Leupold 972h⁻ strain[9] have been sequenced by the Wellcome Trust Sanger Institute and the 13 other laboratories that make up the *S. pombe* European Sequencing Consortium (EUPOM), together with 100 kb of sequence generated by the Cold Spring Harbor Laboratory (GenBank accession numbers AL355920, AL355921, AL391034 and AL391016). Here, we present and discuss the genome sequence and composition, and carry out an initial overview of gene function, making comparisons with other eukaryotic organisms, particularly *S. cerevisiae*.

Mapping, sequencing and sequence analysis

A clone map was generated by the integration of the two pre-existing maps[17,18]. End sequencing and restriction digestion of cosmids were used to construct a minimal tile path for sequencing. Problems with the earlier maps included the existence of chimaeric clones, mismapped cosmids, bacterial insertion elements and unfilled gaps. Small gaps were covered using a long-range polymerase chain reaction (PCR) strategy, plasmid libraries, and a bacterial artificial chromosome (BAC) library provided clones for gap closure across regions not represented in the cosmid libraries. The final 12.5-Mb sequence of the *S. pombe* genome is a composite of 452 cosmids, 22 plasmids, 15 BAC clones and 13 PCR products.

Most sequencing was performed using random sequencing of sub-cloned DNA followed by directed sequencing[19]. DNA from clones was shattered (usually by sonication) and fragments of 1.4–2 kb were cloned, typically, into M13 or pUC18. Random subclones were sequenced with dye-terminator chemistry and analysed on automated sequencers. Most laboratories used Phred software for sequence base calling and Phrap or Gap4 for contig assembly[20]. Gaps and low-quality regions of the sequence were resolved using primer walking, PCR and re-sequencing clones, under conditions that gave increased read lengths. Some laboratories also used direct blotting procedures, classical radioactive sequencing and nested deletions. All sequences were finished to a high degree of accuracy, with at least two high-quality reads on each strand, or, if this could not be accomplished, an additional read on the same strand using an alternative chemistry. The depth of coverage was on average eight-fold. Sequences were collected centrally at the Wellcome Trust Sanger Institute, where the quality was examined by comparison of overlapping regions and by checking for frameshifts in coding regions. The sequencing error rate was less than 1 in 180,000 base pairs (bp), calculated from the number of single-base differences observed in overlapping sequences from different sources. All identified sequencing errors have been resolved with the exception of four single-base differences found in homopolymeric tracts located outside coding regions, possibly generated by slippage during DNA replication.

Gene prediction was carried out with GENEFINDER (P. Green and L. Hillier, unpublished software) trained on experimentally confirmed *S. pombe* genes to recognize intronic and coding regions. Additional information was provided using a Hidden Markov Model trained on intron sequences using HMMER (http://hmmer.wustl.edu/hmmer-html/). Searches were performed against public databases (SWISS-PROT and TrEMBL[21], EMBL[22] and Pfam[23]), using BLAST[24], MSPcrunch[25], FASTA[26] and Genewise[27]. The predictions were refined manually within the Artemis analysis and annotation tool[28] using protein homology and expressed sequence tag (EST) data[29]. Because most *S. pombe* genes have a prospective homologue in other organisms, putative functions were assigned on the basis of similarities to known genes, using the SWISS-PROT[21], Pfam[23], Proteome[30], SGD[31] and MIPS databases[32]. Identification of transfer RNA was carried out using the tRNA scan-SE software[33].

Prediction of genes in fission yeast is a problem of intermediate complexity. It is more difficult than the analysis of tightly packed genomes that have little or no splicing, as found in prokaryotes and budding yeast, but less difficult than gene prediction in multicellular eukaryotes, which have lower gene density, high levels of splicing, and long introns. There are 4,730 confirmed and predicted introns in *S. pombe*, many more than the 272 now predicted for *S. cerevisiae*. *S. pombe* introns average only 81 nucleotides in length and so are shorter and easier to predict than those found in Metazoa and plants. Of the 4,730 introns in *S. pombe*, 638 have been confirmed experimentally by messenger RNA and EST data[29], and many more by homology.

Genome content

We predicted a maximum of 4,940 protein coding genes (including 11 mitochondrial genes) and 33 pseudogenes. The three gene maps showing these predictions can be viewed at ftp://ftp.sanger.ac.uk/pub/yeast/pombe/GeneMaps/. All open reading frames (ORFs) over 100 amino acids with an initiator methionine and not overlapping with other known genes are included in this set. Also included are 147 confirmed or predicted protein-coding sequences of 25–99 amino acids. Any remaining undiscovered genes are likely to have either a highly spliced structure with small exons, or to be smaller than 100 amino acids. There are a further 116 questionable proteins considered less likely to be coding because they are small, have no detectable homologies, and display low coding potential. Removal of these questionable genes reduces the predicted gene complement from 4,940 to 4,824.

Even our upper estimate of 4,940 genes for *S. pombe* is substantially less than the 5,570–5,651 genes predicted for *S. cerevisiae*[34,35], the 6,752 genes predicted for *Mesorhizobium loti*, the largest published prokaryote genome sequence to date[36], and the 7,825 genes estimated in the 8.67-Mb genome of the prokaryote *Streptomyces coelicolor* (J. Parkhill and S. Bentley, personal communication). We conclude that a free-living eukaryotic cell can be constructed with fewer than 5,000 genes, and that the distinction between eukaryotic and prokaryotic cell organization is not determined simply by total number of genes but depends on the types of genes present and how they interact with each other and the environment. Comparing the genome content of species at different levels of organization, it seems that fewer than 500 genes are sufficient to generate a parasitic prokaryotic cell such as *Mycoplasma genitalium*[37], about 1,500 genes for a free-living prokaryotic cell such as *Aquifex aeolicus*[38], 5,000 genes for a free-living eukaryotic cell (*S. cerevisiae* and *S. pombe*; ref. 39 and this paper), and around 15,000 genes for multicellular eukaryotic organisms such as *Drosophila* and *C. elegans*[2,3], whereas 30,000–40,000 genes gives rise to human consciousness[5,6].

Gene density is similar for chromosomes I and II, with one gene every 2,483 and 2,457 bp respectively, but is less dense for chromosome III, at one gene every 2,790 bp. This is not due to differences in the average length of the genes, which are similar (1,407–1,446 bp) for all three chromosomes (Table 1). Protein-coding genes are absent from the centromeres, although tRNA genes are found in these regions. Gene density is also lower at the telomeres. The gene density for the complete genome is one gene every 2,528 bp, compared with one gene every 2,088 bp for *S. cerevisiae*. The protein-coding sequence is predicted to occupy 60.2% (57% excluding introns) of the sequenced portion of the *S. pombe* genome, compared with 71% in *S. cerevisiae* (70.5% excluding introns). The overall guanine and cytosine (GC) content is 36.0%, compared with 38.3% in *S. cerevisiae*, and for the protein-coding portion is identical in the two yeasts at 39.6%.

We have identified a total of 174 tRNAs, 45 of which have introns; all the tRNA families needed to decode all codons are present. The spliceosomal RNAs (U1–U6) are found together with 16 small nuclear RNA genes (snRNAs) and 33 small nucleolar RNAs (snoRNAs). These are dispersed mostly as singletons throughout the genome. The 5.8S, 18S and 26S ribosomal RNA genes are grouped

© 2002 Macmillan Magazines Ltd

Table 1 Genome content for the three chromosomes

	Length (bp)	No. of genes	No. of Tf2s	No. of pseudo Tf2s	No. of wtfs	No. of lone LTRs	No. of pseudogenes	Mean gene length (bp)*	Gene density†	Coding (%)
Chromosome 1	5,598,923	2,255	8	0	1	77	17	1,446	2,483	58.6
Chromosome 2	4,397,795	1,790	2	1	1	53	9	1,411	2,457	57.5
Chromosome 3	2,465,919	884	1	2	23	50	7	1,407	2,790	54.5
Whole genome	12,462,637	4,929	11	3	25	180	33	1,426	2,528	57.5

* Mean gene length excluding introns.
† Gene density, given as average bp per gene.

together as 100–120 tandem repeats in two arrays on chromosome III[40], but the thirty 5S ribosomal RNA genes are distributed throughout the genome[41], providing opportunities for unequal crossing over when they are in tandem orientation and close proximity. This can lead to local duplications and deletions of genes located between the 5S RNA genes[42]. There are 11 intact transposable elements (Tf2 type) (Table 1), accounting for 0.35% of the genome. This is significantly less than the 2.4% (59 elements) found in *S. cerevisiae*[43] and the 10% found in *Arabidopsis*[4], and is also likely to be much less than the numbers in *Drosophila* and humans[44,45]. There are 25 wtf elements ('with tf1- or tf2-type' long terminal repeats, LTRs), which appear to be spliced membrane proteins of *S. pombe*. These elements are often flanked by LTRs, and so may have been duplicated by retrotransposition. There are also 180 solo LTRs, marking former transposition events, compared with 268 found in *S. cerevisiae*. The density of transposable element remnants on chromosome III of *S. pombe* is twice that of chromosomes I and II (Table 1).

We examined 73 genetically and physically mapped genes from the three gene maps; comparison of these maps shows that they are essentially co-linear and that the level of recombination is similar throughout the three chromosomes. More detailed comparisons of the genetic and physical maps may reveal subtle variations in recombination around centromeres, telomeres, the mating-type locus, and sites of meiotic DNA double-strand breaks. Several inconsistencies in the genetic maps were identified, including the reversal of a chromosome II fragment near the telomere between *trp1* and *spo4* (ref. 46), the relocation of *cut1* and *wee1* from the telomere region to the centromere region of chromosome III, and changes in position of *lys1* and *top1*.

Centromere structures

The outline structure of the centromeres has previously been deduced by Southern blotting and by sequencing about 14% of the centromere repeat regions[47–49]. Here, we sequenced most (81%) of the three centromeres; this has allowed schematic maps of the centromeres to be verified (Fig. 1). The nomenclature used follows that of the Yanagida group[50,51]; however, other designations of the centromere elements have been used[52]. The most complete sequence is for centromere 1, which is the shortest at 35 kb and is missing only one 2.5-kb fragment. This centromere consists of a central core (cnt1) of 4.1 kb and 28% GC content, flanked by two 5.6-kb imperfect imr1 repeats (imr1L, imr1R) with 29% GC content, and two pairs of 4.4-kb dg and 4.8-kb dh repeats (dg1, dh1) of 33–34% GC content. A repeat of around 0.3 kb, known as cen 253 (EMBL X13757), is found adjacent to the dh repeats. The maps of the other two centromeres have the same basic structure with central cnt regions flanked by imr repeats and by variable numbers of dg and dh repeats separated by cen 253. Cnt1, -2 and -3 share 48% identity over a 1,405-bp region, and dh1, -2 and -3 share 48% identity over a 1,811-bp region. However, the most striking conservation is observed in the dg regions, which share 97% identity over a 1,780-bp region. This highly conserved segment represents an element that is essential for centromere function; deletion of this

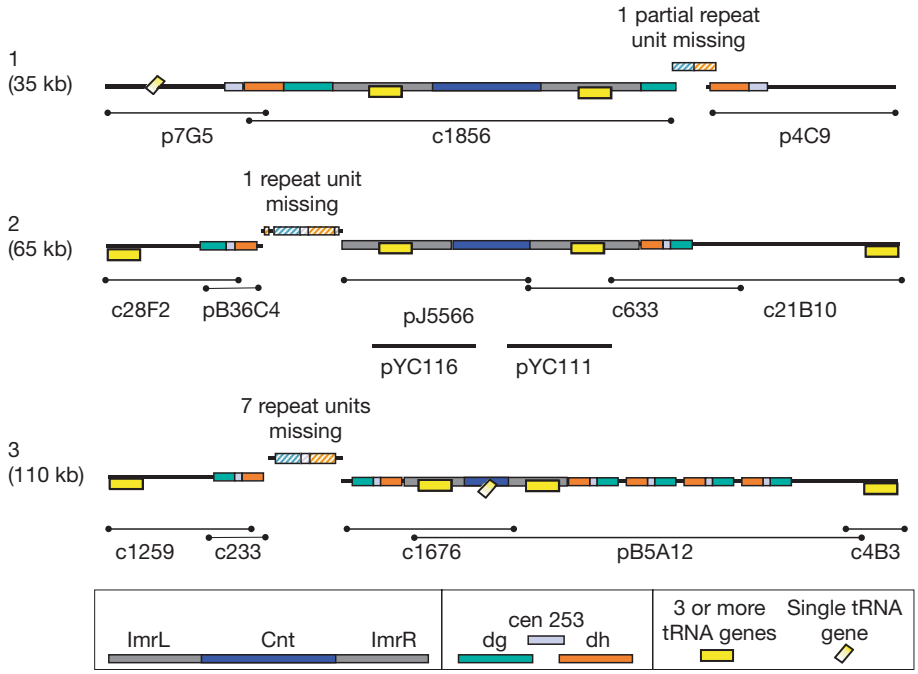

Figure 1 Schematic maps of the three *S. pombe* centromeres showing the repeated elements. The key is given at the bottom of the figure and the relevant clones are indicated under each centromere map. The maps are not drawn to scale.

© 2002 Macmillan Magazines Ltd

region from the dg repeat, termed the K/K″ repeat by the Clarke group, results in a complete loss of centromere activity in both mitosis and meiosis[53]. There must be a special mechanism to maintain such a high level of sequence conservation between the different centromeres. The total calculated lengths of centromeres 1, 2 and 3 are respectively 35, 65 and 110 kb, inversely proportional to the lengths of the chromosomes at 5.7, 4.6 and 3.5 Mb. Possibly more extended centromeric regions are required for proper mitotic and meiotic behaviour when the chromosome arms are shorter. As noted above there are no protein-coding genes in the centromeric region but there are many tRNA genes (Fig. 1). tRNA clusters flank centromeres 2 and 3 and are also found within the imr regions of all three centromeres[50]. These tRNA genes might contribute to centromere function by defining domain boundaries important for centromere activity[54].

The *S. pombe* centromeres are considerably longer than their *S. cerevisiae* equivalents, which contain a core region sufficient for centromere activity of only 120 bp[55,56] and a nuclease-protected region of 150–160 bp including the 120-bp conserved core[57]. It is not clear why *S. pombe* centromeres are 300–1,000 times larger than their *S. cerevisiae* equivalents, but one possibility is that their kinetochore structures are different.

Intergene regions

The total intergene length distributions for *S. pombe* and *S. cerevisiae* are shown in Fig. 2. The length is calculated from the stop codon to the next start codon for tandemly oriented genes, from the start codon to the start codon for divergently oriented genes, and from the stop codon to the stop codon for convergently oriented genes. Intergenic regions in *S. pombe* have a mode of 423 bp and a mean of 952 bp, both longer than the equivalent values for *S. cerevisiae* (200 and 515 bp respectively). Analysis of the divergent intergene regions reveals that pairs of upstream regions range in length from 200 to 2,100 bp, with a peak between 200 and 1,200 bp (Fig. 2). This is longer than the equivalent distributions in *S. cerevisiae*, which range from 200 to 900 bp, with a peak from 200 to 700 bp (Fig. 2). Analysis of convergent intergene regions shows a peak in length for pairs of downstream regions of 200–800 bp for *S. pombe* and 100–500 bp for *S. cerevisiae* (Fig. 2). Therefore there is a smaller difference between the two yeasts for the intergenic regions between convergent genes (downstream regions) than for those between the divergent genes (upstream regions).

Several explanations can account for these results. The 5′ mRNA regions may be systematically longer in *S. pombe* than in *S. cerevisiae*, although there is no evidence for this. For example, the spacing between the TATA-box region and the transcriptional start in *S. pombe* is shorter than that in *S. cerevisiae*[58,59]. Alternatively, the promoter regions may be of greater complexity in *S. pombe* and therefore longer. Again there is no direct evidence to support this view, but there are other examples of more-extended organization of chromatin elements in *S. pombe*, including larger centromeres and regions of DNA replication origin[60]. The existence of truly intergenic spacer regions in *S. pombe* is supported by the identification of several 4–8-kb extended gene-free regions, which fall outside the broad distribution of lengths associated with average intergenic regions. These are low complexity sequences with a $(G-C)/(G+C)$ strand switch[61]. There are about ten gene-free regions per chromosome, which are usually flanked by tandemly oriented genes. One of these gene-free regions, between SPAC4G8.03c and SPAC4G8.04, corresponds to a prominent meiotic DNA break site or cluster of sites (J. A. Young, R. W. Schreckhise and G. R. Smith, manuscript in preparation).

Introns

A total of 4,730 introns is distributed among 43% of *S. pombe* genes, with 15 being the largest number of introns found within a single gene (Table 2). Introns varied from 29 to 819 nucleotides long, with a mean length of 81 and a mode of 48 nucleotides. In *S. cerevisiae*, introns are much rarer, with only 5% of genes having introns. Most introns in *S. pombe* follow the rule of GT donor and AG acceptor, but there are three examples that have GC donors[62]. The average positions of introns within genes were assessed by mapping them with respect to the start and stop codons. This analysis does not take into account any introns in 5′ and 3′ untranslated regions. For the genes with 1–6 introns there is a 5′ bias from the values expected if introns were evenly distributed throughout the genes (Table 2). A 5′ bias is also seen in *S. cerevisiae*, where it has been hypothesized to be due to *in vivo* reverse transcription generating complementary DNAs primed from the 3′ ends of the mRNAs, followed by replacement of the original chromosomal gene with the cDNA by homologous recombination[63]. Because cDNAs are extended from their 3′ ends, there will be a tendency for introns at 5′ ends not to be

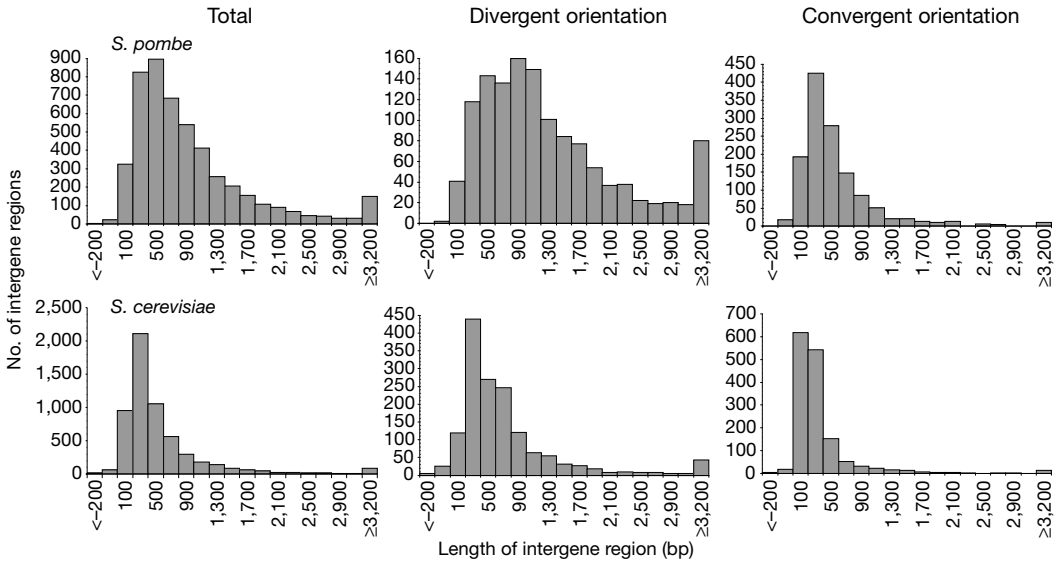

Figure 2 Intergene regions. Distribution of intergene regions given for all genes and for divergent and convergent pairs of genes, for both *S. pombe* and *S. cerevisiae*. A total of 4,890 intergene regions from *S. pombe* were analysed from a database prepared just before completion of the whole genome, and 5,788 intergene regions from *S. cerevisiae* were analysed. Histograms show the number of regions in 200-bp bins.

© 2002 Macmillan Magazines Ltd

Table 2 Introns per gene and average positions of introns within genes

Introns per gene	No. of genes	Mean gene length (bp)	Position of introns*					
			1	2	3	4	5	6
0	2,683	1,497	–	–	–	–	–	–
1	996	1,426	0.26 (0.50)	–	–	–	–	–
2	614	1,396	0.17 (0.33)	0.48 (0.66)	–	–	–	–
3	324	1,588	0.13 (0.25)	0.37 (0.50)	0.63 (0.75)	–	–	–
4	148	1,633	0.10 (0.20)	0.27 (0.40)	0.50 (0.60)	0.73 (0.80)	–	–
5	70	1,603	0.08 (0.17)	0.22 (0.33)	0.37 (0.49)	0.56 (0.66)	0.77 (0.83)	–
6	40	2,162	0.06 (0.14)	0.22 (0.28)	0.34 (0.42)	0.49 (0.57)	0.66 (0.71)	0.82 (0.85)
7–15	34	2,766	–	–	–	–	–	–

The data set of 4,677 introns was prepared just before completion of the whole genome sequence.
*The mean position of introns, with the values in brackets representing the position if the introns were distributed evenly throughout the gene.

removed from the chromosomal genes. Of genes that have two or more introns, 614 have two introns, 324 have three, 148 have four, 70 have five and 40 have six (Table 2). Thus the number of genes having an extra intron decreases by about half as intron number increases from two to six per gene. These observations may be of relevance to speculations concerning the mechanisms by which introns are generated and removed[64]. The relatively large number of introns in *S. pombe* provides opportunities for alternative splicing to generate protein variants, which could have regulatory roles as well as increasing the range of protein types present in the cell[65].

Genome duplications and comparisons

Comparisons of chromosomal sequences and searches for tracts of conserved gene order did not reveal evidence for large-scale genome duplications in *S. pombe*. This differs from reports for *S. cerevisiae* and *Arabidopsis*, which have suggested that both of these organisms have undergone some large-scale genome duplication[4,66]. However, blocks of duplicated sequence totalling about 50 kb retaining a conserved gene order can be found at the sub-telomeric regions of chromosomes I and II. Twenty-four genes (in groups of two or four) are 100% identical at the DNA level, and twenty of these are localized in sub-telomeric regions, suggesting frequent exchange of genetic information at these positions. Most of these genes code for proteins belonging to families specific to fission yeast and are predicted to be cell-surface proteins. Interestingly, in *S. cerevisiae* 7 of the 16 genes (in groups of two, three or four) that are 100% identical at the DNA level are also located in sub-telomeric regions. These gene products include members of the budding-yeast-specific PAU and COS families, which are also predicted to be cell-surface proteins[39]. In the highly plastic telomeric and sub-telomeric regions of malaria and several other protozoan parasites, genes coding for species-specific cell-surface proteins are also found, for example, the Var, Rifin and Stevor families of *Plasmodium falciparum*[67]. These data suggest that recombination events between telomeric regions may be a major mechanism involved in the generation of organism-specific cell-surface molecules. These molecules may also be of importance for cell identity and for processes that generate hypervariable cell-surface molecules relevant for self and non-self recognition.

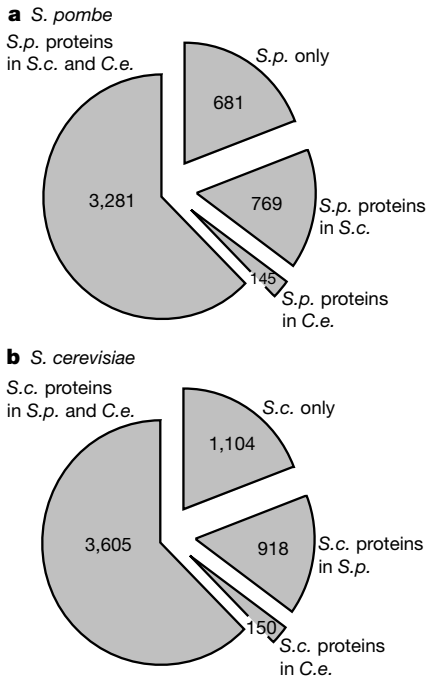

Figure 3 Comparison of proteins in *S. pombe* (*S.p.*), *S. cerevisiae* (*S.c.*) and *C. elegans* (*C.e.*). **a**, Pie chart comparing the homology of proteins of *S. pombe* with those of *S. cerevisiae* and *C. elegans*. **b**, Pie chart comparing the homology of proteins of *S. cerevisiae* with those of *S. pombe* and *C. elegans*. For example, *S.p.* proteins in *S.c.* and *C.e.* means *S. pombe* proteins with homologues found in *S. cerevisiae* and *C. elegans*. The absolute numbers of proteins are given for both yeasts.

We next compared the proteins of *S. pombe* with those of the unicellular eukaryote *S. cerevisiae* and the metazoan *C. elegans* (Fig. 3), using BlastP[24] with a cutoff *E*-value of 0.001 and no low-complexity filtering. Excluding genes coded by the mitochondria and transposons, we used a data set of 4,876 proteins from *S. pombe*, 5,777 proteins from *S. cerevisiae* (Cerpep 14 May 2001; ftp://ftp.sanger.ac.uk/pub/yeast/SCreannotation/cerpep) and 19,622 proteins from *C. elegans* (ftp://ftp.sanger.ac.uk/pub/databases/wormpep). About two-thirds of the *S. pombe* proteins (3,281) have homologues in common with both *S. cerevisiae* and *C. elegans* (Fig. 3). A smaller number, 769 (16%), have homologues in *S. cerevisiae* but not in *C. elegans* and many fewer, 145 (3%), have homologues in *C. elegans* but not in *S. cerevisiae*. A total of 681 proteins (14%) seems to be unique to *S. pombe*. A comparison between *S. cerevisiae* and the other two organisms gave similar results, with 3,605 (62%) of the proteins in common, 918 (16%) found only in *S. pombe* and 150 (3%) only in *C. elegans*, leaving 1,104 proteins (19%) unique to *S. cerevisiae*. Thus, *S. cerevisiae* proteins with homologues only in *S. pombe* total 918 whereas the reverse comparison totals 769 (Fig. 3), indicating that there might be more gene duplications in *S. cerevisiae*, accounting for the extra proteins found in this organism.

To investigate gene duplication further, we carried out an 'all against all' comparison using the same protein data sets and NCBI BlastClust[68] (ftp://ncbi.nlm.nih.gov/blast/documents/README.bcl) to distinguish protein clusters from proteins represented uniquely. Of the 4,876 protein-coding genes of *S. pombe*, 4,515 have no other sequence relatives within the organism and can be considered unique. The remaining 361 are distributed among protein cluster groups with two or more members (Table 3). Using the same parameters in *S. cerevisiae*, 5,061 genes are unique and 716 fall into groups with two or more members (Table 3). This supports the idea that there is less gene redundancy than in *S. cerevisiae*, which may help functional analyses of those genes that are not duplicated in *S. pombe*.

© 2002 Macmillan Magazines Ltd

Table 3 Gene duplication in *S. pombe* and *S. cerevisiae*

Protein members per cluster	No. of clusters in *S. pombe*	No. of clusters in *S. cerevisiae*
1	4,515	5,061
2	124	256
3	17	28
4	8	11
5	2	3
6	1	1
7	2	1
>7	0	3
Total no. of clusters	4,669	5,364
Total no. of sequences	4,876	5,777

Protein clusters were identified with NCBI BlastClust using parameters S10.L0.9, as recommended by Y. Wolf (personal communication). We used databases of 4,876 *S. pombe* proteins prepared just before completion of the genome sequence and of 5,777 *S. cerevisiae* proteins.

Human disease genes

To assess the usefulness of *S. pombe* for investigating the functions of genes related to human disease, we used the same method and dataset of human disease genes as that employed for analysis of the *Drosophila* genome[69]. Protein-coding genes of *S. pombe* were identified that generate products with similarities to proteins coded by 289 genes that are mutated, amplified or deleted in human disease. A total of 172 *S. pombe* proteins have similarity with members of this data set of human disease proteins, and 122 of these have *E*-values greater than 1×10^{-40}. These values indicate that either they are not significant or they have only limited similarities with the equivalent human proteins, reflecting, for example, shared domains such as related protein-interacting regions or catalytic sites. However, despite this limitation, they may still be useful for investigating the biochemical activities and interactions of human disease proteins in *S. pombe*. The other 50 *S. pombe* proteins (Tables 4 and 5) have *E*-values lower than 1×10^{-40}. The more significant similarities seen with this class mean that genes coding for these proteins are more likely to be useful for investigating not only the biochemical but also the biological functions of the human genes, and some could provide good models for studying the associated human disease pathways. The largest group of human disease-related genes are those implicated in cancer. There are 23 such genes (Table 4), and they are involved in DNA damage and repair, checkpoint controls, and the cell cycle, all processes involved in maintaining genomic stability. The cell cycle and checkpoint background of *S. pombe* make it a good model organism for studying these particular cancer disease pathways. Other categories that are also represented in *S. pombe* are those involved in metabolic (12 genes), neurological (13 genes), cardiac (1 gene) and renal (1 gene) disease (Table 5).

A similar analysis in *S. cerevisiae* identified 182 proteins with similarities to the human disease set, with most of the genes coding for these proteins being shared by the two yeasts. Only two of the genes (SPAC630.13c and SPBC530.12c), found in *S. pombe* but not *S. cerevisiae*, code for proteins with any significant similarity to human disease proteins. These are tuberous sclerosis 2 (TSC2), involved in cancer, and ceroid lipofuscinosis PPT1, involved in metabolism. Both yeasts seem to be similarly useful as model organisms for the study of human disease gene function, although their differing biologies may favour one organism for certain genes and the other organism for other genes.

Protein domains

Listed in Table 6 are the ten most frequent protein domains found in *S. pombe*, with 11 more domains of interest in the top 40 most frequent, as determined by InterPro matches[70], together with the frequency of these domains for the other fully sequenced eukaryotic genomes. These domains are divided into three categories (1–3).

The first category (1) consists of five domains found in the top ten most frequent domains in *S. pombe* that are also found in the top ten of at least four of the other eukaryotes. They are the ATP/GTP binding site, the WD40 repeat, the eukaryotic protein kinase catalytic core, the RNA binding region RNP-1, and the zinc finger C2H2-type transcriptional activator. These universal and commonly exploited domains also feature highly in other eukaryotes. Because total gene number increases with the complexity of an organism, the proportion of these domains is approximately similar in each of the sequenced eukaryotic genomes. Energy utilization exploiting the ATP/GTP binding site, protein phosphorylation dependent on the catalytic protein kinase domain, and transcriptional activation using the zinc finger C2H2 domain must define biochemical mechanisms that are readily exploited to generate new biological pathways.

In the second category (2), the domains are present in a similar absolute number in the eukaryotic genomes analysed. Amongst those more frequently found in this category are the BRCT, replication factor C, minichromosome maintenance proteins (MCMs), Fizzy, DNA-directed DNA polymerase β family and helicase C-terminal domains. Some of these are involved in core cell activities like DNA replication, DNA repair and cell-cycle progression, perhaps explaining why they are present in similar

Table 4 *Schizosaccharomyces pombe* genes related to human cancer genes

Human cancer gene	Score*	*S. pombe* gene/product	Systematic name
Xeroderma pigmentosum D; *XPD*	++++	rad15, rhp3	SPAC1D4.12
Xeroderma pigmentosum B; *ERCC3*	++++	rad25	SPAC17A5.06
Hereditary non-polyposis colorectal cancer (HNPCC); *MSH2*	++++	msh2	SPBC24C6.12C
Xeroderma pigmentosum F; *XPF*	++++	rad16, rad10, rad20, swi9	SPCC970.01
Immunodeficiency; DNA ligase 1	++++	cdc17	SPAC57A10.13C
HNPCC; *PMS2*	++++	pms1	SPAC19G12.02C
HNPCC; *MSH6*	++++	msh6	SPCC285.16C
HNPCC; *MSH3*	++++	swi4	SPAC8F11.03
HNPCC; *MLH1*	++++	mlh1	SPBC1703.04
Haematological Chediak–Higashi syndrome; *CHS1*	++++	–	SPBC28E12.06C
Darier–White disease; *SERCA*	++++	pgak	SPBC31E1.02C
Bloom syndrome; *BLM*	++++	hus2, rqh1, rad12	SPAC2G11.12
Ataxia telangiectasia; *ATM*	++++	tel1	SPCC23B6.03C
Xeroderma pigmentosum G; *XPG*	+++	rad13	SPBC3E7.08C
Tuberous sclerosis 2; *TSC2*	+++	–	SPAC630.13C
Immune bare lymphocyte; *ABCB3*	+++	–	SPBC9B6.09C
Downregulated in adenoma; *DRA*	+++	–	SPAC869.05C
Diamond–Blackfan anaemia; *RPS19*	+++	rps19	SPBC649.02
Cockayne syndrome I; *CKN1*	+++	–	SPBC577.09
RAS	+++	ste5, ras1	SPAC17H9.09C
Cyclin-dependent kinase 4; *CDK4*	+++	cdc2	SPBC11B10.09
CHK2 protein kinase	+++	cds1	SPCC18B5.11C
AKT2	+++	pck2, sts6, pkc1	SPBC12D12.04C

*Scores are: ++++, $<1\times10^{-100}$; +++, 1×10^{-40} to 1×10^{-100}.

© 2002 Macmillan Magazines Ltd

Table 5 *Schizosaccharomyces pombe* genes related to human disease genes

Human disease gene	Disease	Score*	*S. pombe* gene/product	Systematic name
Wilson disease; *ATP7B*	Metabolic	++++	P-type copper ATPase	SPBC29A3.01
Non-insulin-dependent diabetes; *PCSK1*	Metabolic	++++	krp1, kinesin related	SPAC22E12.09C
Hyperinsulinism; *ABCC8*	Metabolic	++++	ABC transporter	SPAC3F10.11C
G6PD deficiency; *G6PD*	Metabolic	++++	zwf1 GP6 dehydrogenase	SPAC3A12.18
Citrullinaemia type I; *ASS*	Metabolic	++++	Argininosuccinate synthase	SPBC428.05C
Wernicke–Korsakoff syndrome; *TKT*	Metabolic	+++	Transketolase	SPBC2G5.05
Variegate porphyria; *PPOX*	Metabolic	+++	Protoporphyrinogen oxidase	SPAC1F5.07C
Maturity-onset diabetes of the young (MODY2); *GCK*	Metabolic	+++	hxk1, hexokinase	SPAC24H6.04
Gitelman's syndrome; *SLC12A3*	Metabolic	+++	CCC Na-K-Cl transporter	SPBC18H10.16
Cystinuria type 1; *SLC3A1*	Metabolic	+++	α-glucosidase	SPBC1683.07
Cystic fibrosis; *ABCC7*	Metabolic	+++	ABC transporter	SPBC359.05
Bartter's syndrome; *SLC12A1*	Metabolic	+++	CCC Na-K-Cl transporter	SPBC18H10.16
Menkes syndrome; *ATP7A*	Neurological	++++	P-type copper ATPase	SPBC29A3.01
Deafness, hereditary; *MYO15*	Neurological	++++	myo51 class V myosin	SPBC2D10.14C
Zellweger syndrome; *PEX1*	Neurological	+++	AAA-family ATPase	SPCC553.03
Thomsen disease; *CLCN1*	Neurological	+++	ClC chloride channel	SPBC19C7.11
Spinocerebellar ataxia type 6 (SCA6); *CACNA1A*	Neurological	+++	VIC sodium channel	SPAC6F6.01
Myotonic dystrophy; *DM1*	Neurological	+++	orb6 Ser/Thr protein kinase	SPAC821.12
McCune–Albright syndrome; *GNAS1*	Neurological	+++	gpa1 guanine nucleotide binding	SPBC24C6.06
Lowe's oculocerebrorenal syndrome; *OCRL*	Neurological	+++	PIP phosphatase	SPBC2G2.02
Dents; *CLCN5*	Neurological	+++	ClC chloride channel	SPBC19C7.11
Coffin–Lowry; *RPS6KA3*	Neurological	+++	Ser/Thr protein kinase	SPCC24B10.07
Angelman; *UBE3A*	Neurological	+++	Ubiquitin-protein ligase	SPBP8B7.27
Amyotrophic lateral sclerosis; *SOD1*	Neurological	+++	sod1, superoxide dismutase	SPAC821.10C
Oguchi type 2; *RHKIN*	Neurological	+++	Ser/Thr protein kinase	SPCC24B10.07
Familial cardiac myopathy; *MYH7*	Cardiac	++++	myo2, myosin II	SPCC645.05C
Renal tubular acidosis; *ATP6B1*	Renal	++++	V-type ATPase	SPAC637.05C

*Scores are: ++++, $<1\times10^{-100}$; +++, 1×10^{-40} to 1×10^{-100}.

absolute number regardless of genome size[71]. Systematic searches for other domains present in similar absolute numbers in genomes of all eukaryotes might identify other, at present unrecognized, functions involved in similar core cell activities.

The third category (3) includes domains whose occurrence rises dramatically with increasing genome size within the Metazoa. This category includes the SH3, PH and tyrosine/dual-specificity phosphatase domains. These are involved in intra- and intercellular signalling pathways, which might be expected to become increasingly elaborate as multicellular complexity increases[69,71].

Two other domains in the top ten for both the yeasts are the sugar and ABC transporters (Table 6). *S. cerevisiae* has significantly more of these domains and the amino-acid permease domain than does *S. pombe*[72], which may explain why it is a more versatile organism, growing on a greater range of media. The Zn(II)Cys(6) transcription-factor domain is found only in the two yeasts, supporting the idea that it is specific to fungi. The chromodomain is found more frequently in *S. pombe*—seven examples compared with two in *S. cerevisiae*—possibly reflecting differences in higher-order chromatin structure.

Defining the eukaryotic cell

The genome sequence of *S. pombe* increases the range of available complete eukaryotic genome sequences to two unicellular free-living organisms (*S. cerevisiae* and *S. pombe*), one plant (*Arabidopsis*), and three metazoans (*C. elegans*, *Drosophila* and humans). This range of organisms allows a comparison between eukaryotic and prokaryotic genomes (represented by 37 bacteria and 8 archaea), with the intention of identifying those genes important for eukaryotic cell organization. We have made an

Table 6 Protein domain analysis and comparison with other eukaryotes

Interpro accession no.	*S. pombe*		*S. cerevisiae*		*H. sapiens*		*D. melanogaster*		*C. elegans*		*A. thaliana*		Interpro name	
	Proteins	Rank	Proteins	Rank	Proteins	Rank	Proteins	Rank	Proteins	Rank	Proteins	Rank		
IPR001687	213	1	267	1	436	5	231	4	191	7	331	5	ATP/GTP-binding site motif A (Ploop)	1
IPR001680	114	2	97	3	277	8	183	5	102	19	210	10	G protein β WD40 repeats	1
IPR000719	111	3	119	2	579	3	377	2	450	2	1,049	1	Eukaryotic protein kinase	1
IPR000504	80	4	61	5	307	7	182	6	97	21	255	8	RNA binding region RNP1	1
IPR001650	67	5	63	4	155	20	101	17	80	27	148	13	Helicase C-terminal domain	2
IPR001841	44	6	33	12	215	15	120	11	126	12	379	4	RING finger	–
IPR001440	38	7	33	12	150	21	92	18	46	43	125	17	TPR repeat	–
IPR001066	36	8	46	8	44	64	45	34	55	37	98	26	Sugar transporter	–
IPR001617	33	9	42	9	75	40	67	28	61	36	103	25	ABC transporter family	–
IPR000822	32	10	51	7	712	2	403	1	154	10	115	20	Zinc finger, C2H2 type	1
IPR001357	14	23	10	30	24	82	17	61	25	60	17	83	BRCT domain	2
IPR000862	8	29	9	31	8	98	9	68	6	79	13	87	Replication factor C conserved domain	2
IPR002064	5	32	5	35	4	102	6	70	3	82	5	95	DNA directed DNA polymerase family β	2
IPR001208	6	31	6	34	12	94	13	64	5	80	8	92	MCM family	2
IPR000002	5	32	3	37	3	103	4	72	2	83	6	94	FIZZY/CDC20 domain	2
IPR001452	21	16	23	18	220	14	82	23	62	35	3	97	Src homology 3 (SH3) domain	3
IPR001849	21	16	26	16	253	11	89	22	75	31	27	73	PH domain	3
IPR000387	9	28	11	29	112	29	47	40	110	16	21	79	Tyrosine-specific protein phosphatase and dual-specificity protein phosphatase family	3
IPR001138	27	13	52	6	0	NA	0	NA	0	NA	0	NA	Fungal transcriptional regulatory protein	–
IPR002293	21	16	32	13	43	65	36	45	32	54	65	42	Permease for amino acids and related compounds	–
IPR000953	7	30	2	38	26	80	20	58	15	70	24	76	Chromodomain	–

Domain identifiers are from InterPro, which integrates PROSITE, PRINTS and PFAM. Only domains within the most frequent 40 found in *S. pombe* are given. The numbers of proteins with these domains and their ranking is given for *S. pombe* and the other eukaryotes listed. At the right end of the table is a classification of 1–3; see text for an explanation. NA, not applicable.

© 2002 Macmillan Magazines Ltd

Table 7 Identifying conserved genes important for defining the eukaryotic cell and multicellularity

Similarity	No. of genes	≥20%	≥15%	≥12%
(**a**) Genes defining eukaryotic organization				
50%	184	62	47	41
45%	245	86	63	55
40%	311	113	81	70
(**b**) Genes defining multicellularity				
50%	397	1	1	1
45%	511	2	1	1
40%	647	3	2	2

The same data set for assessing gene duplication was used. Protein data sets were identified with 40%, 45% and 50% similarity for humans, *Drosophila*, *C. elegans*, *S. pombe* and *S. cerevisiae* in **a** or for human, *Drosophila*, *C. elegans* and *Arabidopsis* in **b**. The Blast-calculated bit score describes the similarity between two sequences. For two identical sequences (a compared to a) the bit score is 100%. For different sequences (a compared with b) the measure of similarity is bit score (ab)/bit score (aa) × 100. The numbers within these data sets are not found in any of the fully sequenced prokaryotes (45 in total) in **a**, or any of the prokaryotes and the two yeasts in **b** at similarity levels of 12%, 15% and 20%. The 45 prokaryotes include genomes from 37 Eubacteria and 8 Archaea.

initial analysis to identify the more conserved genes falling in this category by comparing the predicted protein sequences coded by the above genomes. The percentage similarity was derived from the hit bit score divided by the self bit score for each protein (see Table 7 legend). We selected those proteins with a high percentage similarity score in all of the eukaryotes, and a low one in all of the prokaryotes. Three thresholds (50%, 45% and 40%) were used to identify proteins that are highly conserved in the fully sequenced eukaryotes and three corresponding thresholds (20%, 15% and 12% respectively) to identify proteins not found in the fully sequenced prokaryotes (Table 7a). For an initial discussion of these proteins, thresholds of 50% and 20% were selected. This analysis identifies genes coding for proteins that are highly conserved in yeasts, plants and metazoans (by using a threshold of 50% similarity) and yet are not well conserved in prokaryotes (by using a threshold of 20% similarity). The proteins identified using these criteria are likely to be important for maintaining eukaryotic cell organization, although the high threshold of 50% means that other proteins required for this may well be excluded.

Using these thresholds, 62 genes were identified and grouped according to function (Table 8). More information about these genes can be found on the GeneDB website (http://www.genedb.org/pombe) and the PombePD website (http://proteome.com/databases). Two of these groups code for proteins associated with characteristics considered to distinguish eukaryotic cells from prokaryotic cells: the organization of DNA in chromosomes within a nucleus, and the formation of 40S and 60S ribosomal subunits, which are larger than the prokaryotic 30S and 50S subunits. The first group includes the H3 and H4 core histone proteins required for packaging DNA into nucleosomes, the Hda1 histone deacetylase, which suggests histone acetylation is critical for eukaryotic chromatin, and the Ran GTPase Spi1, a key element for nuclear membrane transport. One putative protein in this category (SPAC890.07c) is possibly involved in export of mRNA binding proteins and another may be localized in the nucleus (SPCP1E11.08). The second group includes two Rps and six Rpl proteins, components of the 40S and 60S ribosomal subunits respectively; these eight proteins may contribute to differences in protein translation between prokaryotes and eukaryotes.

Two further groups in Table 8 are relevant for the more elaborate organization and compartmentation of eukaryotic cells. One consists of cytoskeletal proteins, the actins Act1 and Act2, the tubulins Nda2, Nda3 and Tub1, and the cytoskeleton-associated proteins Arp2 and Cdc42. The actin and tubulin polymers provide not only internal structure but also the means for transport of components and information from one region of the cell to another, important matters given the increased size of eukaryotic cells. The bacterial FtsA, Hsp70 and FtsZ proteins have structures with similarities respectively to actin and tubulin but only very limited primary sequence similarities[73–75]. Arp2 is an actin-related protein required for actin organization, and the Cdc42 GTPase is a signalling molecule important for cell shape and for communicating signals from the cytoskeleton. One protein (SPAC926.07c) is predicted to be a dynein light chain. The second group consists of GTP binding proteins and their regulators Ypt1, -2, -3 and -7, Arf1, Aps1, Gdi1 and Sar1, which are required for membrane transport. Membrane-bound organelles and structures are characteristic features of eukaryotic cells, and membrane fusion and fragmentation are important in organelle formation and function. Cam1 (calmodulin) is a protein that exploits compartmentalization of Ca^{2+} to regulate cellular processes. One protein (SPBC1539.08) is a putative ADP ribosylation factor and may be involved in transport.

A small group (Table 8) includes cell-cycle and checkpoint control proteins. The Cdc2 protein kinase (Cdc28 in *S. cerevisiae*) is a cyclin-dependent kinase (CDK) controlling the onset of S-phase and mitosis in the two yeasts, with closely related CDKs controlling these cell-cycle transitions in other eukaryotes. The CDK system for cell-cycle control evolved with the appearance of eukaryotic cells, whose cell cycle differs from prokaryotes in two ways: DNA synthesis, which uses multiple origins of replication, and mitosis, which brings about chromosome segregation. It has been argued that, in the primeval eukaryote, there was a single CDK that underwent a monotonic change during the cell cycle, initiating S phase early in the cycle at a low activity and mitosis late in the cycle at a high activity[76]. Two checkpoint proteins, Rad24 and Rad25, are 14-3-3 proteins thought to regulate the Cdc25 phosphatase controlling the Cdc2 CDK[77]. If DNA becomes damaged then these checkpoint proteins prevent the onset of mitosis until the damage is repaired. This pathway is essential for maintaining genomic stability and seems to be characteristic of eukaryotic cells.

Three further groups reflect biochemical processes that are important in eukaryotic cell regulation. The first group consists of Lsm2 and Smd2, which are required for RNA splicing. The second group consists of the Ubc, Ubi and Ubl proteins together with Uip1 and Pad1 (Table 8), all required to bring about controlled proteolysis of proteins. A further protein putatively involved in proteolysis is a prohibitin complex subunit (SPAC1782.06c). The

Table 8 Classification of conserved genes important for defining the eukaryotic cell

Nucleus	Ribosomal	Cytoskeleton	Compartmentation	Cell cycle	Splicing	Proteolysis	Kinase/ phosphatase	Miscellaneous
h3.1	rpl18	act1	ypt1	cdc2	lsm2	ubc13	cka1	SPBC24C6.11
h3.2	rpl27	act2	ypt2	rad24	smd2	ubc4	dis2	SPBP8B7.24C
h3.3	rpl27A	arp2	ypt3	rad25		ubi1	hhp1	
h4.1	rpl29	cdc42	ypt7			ubi4	ppa1	
h4.2	rpl7A	nda2	aps1			ubl1	ppa2	
h4.3	rpl7	nda3	arf1			uep1	ppe1	
hda1	rps3A	tub1	cam1			hus5	sds21	
spi1	rps21	SPAC926.07C	gdi1			pad1	SPBC26H8.05C	
SPAC890.07C			sar1			rhp6	SPAC22H10.04	
SPCP1E11.08			SPBC1539.08			SPAC1782.06C		

The 62 proteins from Table 7a (50% versus 20%) are classified according to their primary function as described in the text. For putative functions, only the gene location is given.

© 2002 Macmillan Magazines Ltd

third group consists of protein kinases and phosphatases, and includes Cka1, Dis2, Hhpt, Ppa1, Ppa2, Ppe1 and Sds21 and putative serine/threonine protein phosphatases (SPAC22H10.04 and SPBC26H8.05c). The presence of these three regulatory processes unique to eukaryotic cells allows protein levels and activities to be specifically and rapidly changed without relying on changes in transcription rate. In prokaryotic cells, gene regulation often operates through changes in transcription rate, followed by dilution of remaining proteins as a consequence of rapid cellular growth. The slower growth rates of eukaryotic cells means that mechanisms in addition to dilution by growth are required to modulate protein activity; these mechanisms may be provided by RNA splicing, proteolysis and phosphorylation.

Two genes code for a putative zinc-finger protein (SPBC24C6.11) with a possible role in cell polarity and a putative autophagy protein (SPBP8B7.24c) that may mediate attachment of autophagosomes to microtubules. Extension of this analysis at different thresholds of similarity should identify further proteins of unknown function that are important for eukaryotic cell organization.

We performed a similar analysis to identify highly conserved genes that may be important for maintaining multicellular eukaryotic organization (Table 7b). We compared the proteins in prokaryotes and in *S. cerevisiae* and *S. pombe*, which are all unicellular, with those of *C. elegans*, *Drosophila*, *Arabidopsis* and humans, which are all multicellular. The same thresholds were used to identify those proteins that are highly conserved in the four multicellular eukaryotes (50%, 45% and 40%) and to identify which of these proteins were not found to be highly conserved in the unicellular organisms (20%, 15% and 12%). The number of genes coding for proteins that fall into these categories was very small: one to three depending on the thresholds used. These genes code a putative transcription factor, an RNA-binding protein and a selenium-binding protein.

As more sequences become available, the groups of genes we have identified as being important for eukaryotic and multicellular organization will inevitably be modified. However, our results allow us to speculate on the evolutionary transitions from prokaryotes to eukaryotes and to multicellularity. The transition to multicellularity may not have required the evolution of many new genes, absent from unicellular organisms. The pathways necessary for multicellular organization could already have been in existence in unicellular eukaryotes. For example, intercellular signalling may have been solved by the sexual needs of primeval, single-celled eukaryotes to seek out and identify an appropriate mating partner. Once signalling between cells had evolved, it could be readily exploited to generate the signalling pathways required for multicellular organization. The highly conserved genes specific to eukaryotes may be necessary for eukaryotic cell organization to be generated. In contrast, the transition from unicellularity to multicellularity may not have required many new genes. Instead it may have used genes already present in unicellular eukaryotes, perhaps by the shuffling of functional domains, to give rise to new combinations, which allowed the development of pathways required for the evolution of multicellularity[2,69,71,78]. If these speculations are correct, they imply that the evolutionary transition from unicellular prokaryotic to unicellular eukaryotic life may have been more complex than the transition to multicellular life. This might provide some explanation as to why it took around 2,300 million years (Myr) to evolve from the first prokaryote to the first eukaryote (thought to have arisen about 3,800 Myr and 1,500 Myr ago, respectively) but only 500 Myr for the evolution of the first multicellular organisms, which arose about 1,000 Myr ago. Further analyses and comparisons should continue to be illuminating about this interesting question of which genes define eukaryotic cells and which define multicellular organisms. □

Received 16 October 2001; accepted 7 January 2002.

1. Goffeau, A. *et al.* The yeast genome directory. *Nature* **387** (suppl.), 1–105 (1997).
2. The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282,** 2012–2018 (1998).
3. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287,** 2185–2195 (2000).
4. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815 (2000).
5. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
6. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
7. Sipiczki, M. Where does fission yeast sit on the tree of life? *Genome Biol.* **1,** 1011.1–1011.4 (2000).
8. Heckman, D. S. *et al.* Molecular evidence for the early colonization of land by fungi and plants. *Science* **293,** 1129–1133 (2001).
9. Leupold, U. Die Verebung von Homothallie und Heterothallie bei *Schizosaccharomyces pombe*. *C.R. Lab. Carlsberg* **24,** 381–475 (1950).
10. Mitchison, J. M. The growth of single cells. I. *Schizosaccaromyces pombe*. *Expl. Cell Res.* **13,** 244–262 (1957).
11. Fantes, P. & Beggs, J. *The Yeast Nucleus* (Oxford Univ. Press, Oxford, 2000).
12. Davis, L. & Smith, G. R. Meiotic recombination and chromosome segregation in *Schizosaccharomyces pombe*. *Proc. Natl Acad. Sci. USA* **98,** 8395–8402 (2001).
13. Humphrey, T. DNA damage and cell cycle control in *Schizosaccharomyces pombe*. *Mutat. Res.* **451,** 211–226 (2000).
14. Smith, C. L. *et al.* An electrophoretic karyotype for *Schizosaccharomyces pombe* by pulsed field gel electrophoresis. *Nucleic Acids Res.* **15,** 4481–4491 (1987).
15. Lang, B. F., Cedergren, R. & Gray, M. W. The mitochondrial genome of the fission yeast, *Schizosaccharomyces pombe*. Sequence of the large-subunit ribosomal RNA gene, comparison of potential secondary structure in fungal mitochondrial large-subunit rRNAs and evolutionary considerations. *Eur. J. Biochem.* **169,** 527–537 (1987).
16. Schaak, J., Mao, J. & Soll, D. The 5.8S RNA gene sequence and the ribosomal repeat of *Schizosaccharomyces pombe*. *Nucleic Acids Res.* **10,** 2851–2864 (1982).
17. Hoheisel, J. D. *et al.* High resolution cosmid and P1 maps spanning the 14 Mb genome of the fission yeast *S. pombe*. *Cell* **73,** 109–120 (1993).
18. Mizukami, T. *et al.* A 13 kb resolution cosmid map of the 14 Mb fission yeast genome by nonrandom sequence-tagged site mapping. *Cell* **73,** 121–132 (1993).
19. Harris, D. & Murphy, L. in *Genomics Protocols* (eds Starkey, M. & Elaswarapu, R.) 217–234 (Humana, Tokawa, New Jersey, 2001).
20. Bonfield, J. K., Smith, K. & Staden, R. A new DNA sequence assembly program. *Nucleic Acids Res.* **23,** 4992–4999 (1995).
21. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res.* **27,** 49–54 (1999).
22. Stoesser, G., Tuli, M. A., Lopez, R. & Sterk, P. The EMBL nucleotide sequence database. *Nucleic Acids Res.* **27,** 18–24 (1999).
23. Bateman, A. *et al.* Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res.* **27,** 260–262 (1999).
24. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
25. Sonnhammer, E. L. & Durbin, R. A workbench for large-scale sequence homology analysis. *Comput. Appl. Biosci.* **10,** 301–307 (1994).
26. Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA* **85,** 2444–2448 (1988).
27. Birney, E., Thompson, J. D. & Gibson, T. J. PairWise and SearchWise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. *Nucleic Acids Res.* **24,** 2730–2739 (1996).
28. Rutherford, K. *et al.* Artemis: sequence visualization and annotation. *Bioinformatics* **16,** 944–945 (2000).
29. Morimyo, M. *et al.* in *Biodefence Mechanisms against Environmental Stress* (eds Ozawa, T., Hori, T. & Tatsumi, K.) 115–123 (Kondansa, Tokyo & Springer, Heidelberg, 1998).
30. Costanzo, M. 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.* **28,** 73–76 (2000).
31. Cherry, J. M. *et al.* SGD: *Saccharomyces* genome database. *Nucleic Acids Res.* **26,** 73–79 (1998).
32. Mewes, H. W. *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **28,** 37–40 (2000).
33. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25,** 955–964 (1997).
34. Blandin, G. *et al.* Genomic exploration of the hemiascomycetous yeasts: 4. The genome of *Saccharomyces cerevisiae* revisited. *FEBS Lett.* **487,** 31–36 (2000).
35. Wood, V., Rutherford, K. M., Ivens, A., Rajandream, M.-A. & Barrell, B. A re-annotation of the *Saccharomyces cerevisiae* genome. *Comp. Funct. Genom.* **2,** 143–154 (2001).
36. Kaneko, T. *et al.* Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti* (supplement). *DNA Res.* **7,** 381–406 (2000).
37. Hutchison, C. A. *et al.* Global transposon mutagenesis and a minimal *Mycoplasma* genome. *Science* **286,** 2165–2169 (1999).
38. Deckert, G. *et al.* The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* **392,** 353–358 (1998).
39. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274,** 546, 563–567 (1996).
40. Barnitz, J. T., Cramer, J. H., Rownd, R. H., Cooley, L. & Soll, D. Arrangement of the ribosomal RNA genes in *Schizosaccharomyces pombe*. *FEBS Lett.* **143,** 129–132 (1982).
41. Mao, J. *et al.* The 5S RNA genes of *Schizosaccharomyces pombe*. *Nucleic Acids Res.* **10,** 487–500 (1982).
42. Carr, A. M., MacNeill, S. A., Hayles, J. & Nurse, P. Molecular cloning and sequence analysis of mutant alleles of the fission yeast cdc2 protein kinase gene: implications for $cdc2^{+}$ protein structure and function. *Mol. Gen. Genet.* **218,** 41–49 (1989).
43. Kim, J. M., Vanguri, S., Boeke, J. D., Gabriel, A. & Voytas, D. F. Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. *Genome Res.* **8,** 464–478 (1998).
44. Ashburner, M. *et al.* An exploration of the sequence of a 2.9-Mb region of the genome of *Drosophila melanogaster*: the *Adh* region. *Genetics* **153,** 179–219 (1999).
45. Gu, Z., Wang, H., Nekrutenko, A. & Li, W. H. Densities, length proportions, and other distributional

© 2002 Macmillan Magazines Ltd

features of repetitive sequences in the human genome estimated from 430 megabases of genomic sequence. *Gene* **259,** 81–88 (2000).
46. Egel, R. Reorientation of the distal region in linkage group IIR of fission yeast. *Curr. Genet.* **24,** 179–180 (1993).
47. Chikashige, Y. *et al.* Composite motifs and repeat symmetry in *S. pombe* centromeres: direct analysis by integration of *Not*I restriction sites. *Cell* **57,** 739–751 (1989).
48. Murakami, S., Matsumoto, T., Niwa, O. & Yanagida, M. Structure of the fission yeast centromere cen3: direct analysis of the reiterated inverted region. *Chromosoma* **101,** 214–221 (1991).
49. Clarke, L. & Baum, M. P. Functional analysis of a centromere from fission yeast: a role for centromere-specific repeated DNA sequences. *Mol. Cell. Biol.* **10,** 1863–1872 (1990).
50. Takahashi, K., Murakami, S., Chikashige, Y., Niwa, O. & Yanagida, M. A large number of tRNA genes are symmetrically located in fission yeast centromeres. *J. Mol. Biol.* **218,** 13–17 (1991).
51. Nakaseko, Y., Adachi, Y., Funahashi, S.-I., Niwa, O. & Yanagida, M. Chromosome walking shows a highly homologous repetitive sequence present on all the centromere regions of fission yeast. *EMBO J.* **5,** 1011–1021 (1986).
52. Fishel, B., Amstutz, H., Baum, M., Carbon, J. & Clarke, L. Structural organization and functional analysis of centromeric DNA in the fission yeast *Schizosaccharomyces pombe*. *Mol. Cell. Biol.* **8,** 754–763 (1988).
53. Baum, M., Ngan, V. K. & Clarke, L. The centromeric K-type repeat and the central core are together sufficient to establish a functional *Schizosaccharomyces pombe* centromere. *Mol. Biol. Cell* **5,** 747–761 (1994).
54. Partridge, J. F., Borgstrom, B. & Allshire, R. C. Distinct protein interaction domains and protein spreading in a complex centromere. *Genes Dev.* **14,** 783–791 (2000).
55. Hyman, A. A. & Sorger, P. K. Structure and function of kinetochores in budding yeast. *Annu. Rev. Cell Dev. Biol.* **11,** 471–495 (1995).
56. Hieter, P. *et al.* Functional selection and analysis of yeast centromeric DNA. *Cell* **42,** 913–921 (1985).
57. Funk, M., Hegemann, J. H. & Philippsen, P. Chromatin digestion with restriction endonucleases reveals 150–160 bp of protected DNA in the centromere of chromosome XIV in *Saccharomyces cerevisiae*. *Mol. Gen. Genet.* **219,** 153–160 (1989).
58. Russell, P. R. Transcription of the triose-phosphate-isomerase gene of *Schizosaccharomyces pombe* initiates from a start point different from that in *Saccharomyces cerevisiae*. *Gene* **40,** 125–130 (1985).
59. Nagawa, F. & Fink, G. R. The relationship between the "TATA" sequence and transcription initiation sites at the *HIS4* gene of *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **82,** 8557–8561 (1985).
60. Gomez, M. & Antequera, F. Organization of DNA replication origins in the fission yeast genome. *EMBO J.* **18,** 5683–5690 (1999).
61. Lobry, J. R. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.* **13,** 660–665 (1996).
62. Maniatis, T. & Reed, R. The role of small nuclear ribonucleoprotein particles in pre-mRNA splicing. *Nature* **325,** 673–678 (1987).
63. Fink, G. R. Pseudogenes in yeast? *Cell* **49,** 5–6 (1987).
64. Robertson, H. M. The large *srh* family of chemoreceptor genes in *Caenorhabditis* nematodes reveals processes of genome evolution involving large duplications and deletions and intron gains and losses. *Genome Res.* **10,** 192–203 (2000).
65. Graveley, B. R. Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.* **17,** 100–107 (2001).
66. Wolfe, K. H. & Shields, D. C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387,** 708–713 (1997).
67. Bowman, S. *et al.* The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature* **400,** 532–538 (1999).
68. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
69. Rubin, G. M. *et al.* Comparative genomics of the eukaryotes. *Science* **287,** 2204–2215 (2000).
70. Apweiler, R. *et al.* The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **29,** 37–40 (2001).
71. Chervitz, S. A. *et al.* Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science* **282,** 2022–2028 (1998).
72. Hunt, C. *et al.* Subtelomeric sequence from the right arm of *Schizosaccharomyces pombe* chromosome I contains seven permease genes. *Yeast* **18,** 355–361 (2001).
73. Kabsch, W. & Holmes, K. C. The actin fold. *FASEB J.* **9,** 167–174 (1995).
74. Itoh, T., Matsuda, H. & Mori, H. Phylogenetic analysis of the third *hsp70* homolog in *Escherichia coli*; a novel member of the *Hsc66* subfamily and its possible co-chaperone. *DNA Res.* **6,** 299–305 (1999).
75. Erickson, H. P. Atomic structures of tubulin and FtsZ. *Trends Cell Biol.* **8,** 133–137 (1998).
76. Fisher, D. L. & Nurse, P. A single fission yeast mitotic cyclin B $p34^{cdc2}$ kinase promotes both S-phase and mitosis in the absence of G1 cyclins. *EMBO J.* **15,** 850–860 (1996).
77. Lopez-Girona, A., Furnari, B., Mondesert, O. & Russell, P. Nuclear localization of Cdc25 is regulated by DNA damage and a 14-3-3 protein. *Nature* **397,** 172–175 (1999).
78. Lundin, L. G. Gene duplications in early metazoan evolution. *Semin. Cell Dev. Biol.* **10,** 523–530 (1999).

Acknowledgements

We thank the European Commission, the Wellcome Trust and Cancer Research UK for financial support. We also thank all the many people in the fission yeast community for their comments and suggestions at all stages of this project, particularly M. Mitchison and U. Leupold, the founders of fission yeast studies. Cancer Research UK, London Research Institute, comprises Lincoln's Inn Fields and Clare Hall Laboratories of the former Imperial Cancer Research Fund following the merger of the ICRF with the Cancer Research Campaign in February 2002.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.A.R. (e-mail: mar@sanger.ac.uk).

© 2002 Macmillan Magazines Ltd