

# The *C. elegans* genome sequencing project: a beginning

J. Sulston\*, Z. Du†, K. Thomas\*, R. Wilson†, L. Hillier†, R. Staden\*, N. Halloran†, P. Green†, J. Thierry-Mieg‡, L. Qiu†, S. Dear\*, A. Coulson\*, M. Craxton\*, R. Durbin\*, M. Berks\*, M. Metzstein\*, T. Hawkins\*, R. Ainscough\* & R. Waterston†

\*MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

†Department of Genetics, Box 8232, Washington University School of Medicine, 4566 Scott Avenue, St Louis, Missouri 63110, USA

‡CNRS-CRBM et Physique-Mathématique, PO Box 5051, Montpellier 34044, France

The long-term goal of this project is the elucidation of the complete sequence of the *Caenorhabditis elegans* genome. During the first year methods have been developed and a strategy implemented that is amenable to large-scale sequencing. The three cosmids sequenced in this initial phase are surprisingly rich in genes, many of which have mammalian homologues.

THE realization that the human genome can be sequenced in its entirety has stimulated great interest in genome analysis<sup>1,2</sup> and efforts have already begun to construct genetic and physical maps<sup>3</sup> and to improve DNA sequencing methods<sup>4</sup>. In pursuit of this great enterprise, it will be necessary to sequence and analyse the smaller genomes of experimentally tractable organisms which will serve as pilot systems for evaluating technology and also provide information essential for interpreting the human sequence. The genomes of single-celled organisms such as *Escherichia coli* and *Saccharomyces cerevisiae* will reveal features peculiar to the basic functions shared by all living things. Analysis of simple animal genomes intermediate in size and complexity between those of yeast and man will help us to understand the more complex features of mammals.

The genome of the small nematode *C. elegans* is a good candidate for complete sequence analysis<sup>5</sup>. This organism has been used to investigate animal development and behaviour<sup>6,7</sup>. Its small size and short generation time facilitate genetic analysis, and more than 900 loci have now been identified through mutations<sup>8</sup>. Each animal develops with essentially the same pattern of cell divisions, and this entire pattern is known<sup>9-11</sup>. The anatomy is simple, with only 959 somatic cells, and the ultrastructure established—for example, the complete connectivity of its 302 neurons has been determined<sup>12</sup>.

FIG. 1 Physical map of the region where the sequencing project has begun. *a*, Selected overlapping cosmid and lambda clones are represented by the horizontal lines. The cosmids which have been sequenced and reported here are underlined in bold, as well as the additional cosmids underway at present. *b*, The overlapping YAC clones from the region bridge the segments not represented in cosmid clones above. Methods must still be developed to capture sequence efficiently from the spans presently cloned in YACs only. *c*, Six genes and markers known to lie in the regions from genetic and other studies<sup>50</sup>. The genes *unc-32* and *sup-5* are less than 0.1 centimorgans apart on the genetic map of chromosome III. Because of the way the physical map was constructed<sup>16,17</sup>, the physical distance separating these two genes can only be estimated to be more than 150 kb, yielding a ratio of more than

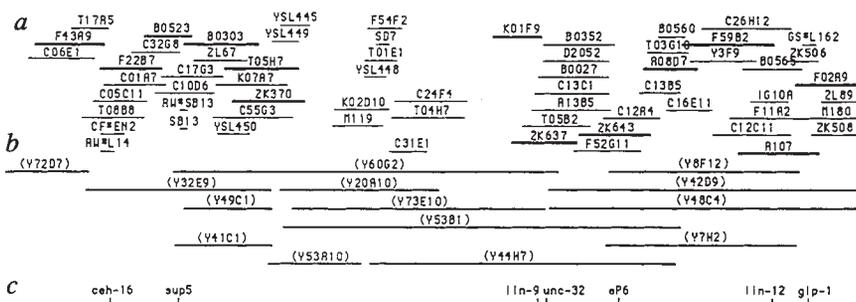
Many genes required for normal development and behaviour are being studied. With the aid of an active transposon system<sup>13-15</sup> and a physical map of the genome<sup>16-18</sup>, they can be readily cloned. Methods for transformation have been developed which allow reintroduction of engineered genes<sup>19-21</sup>. The similarity of many of these genes to those in mammals is often extensive, supporting the contention that information obtained in the nematode will be relevant to understanding the biology of man.

The *C. elegans* genome contains an estimated 100 megabases ( $10^8$  bases), less than the size of an average human chromosome. Generally genes in *C. elegans* have smaller and fewer introns than their mammalian counterparts<sup>22</sup> and the gene density is high (see below). A clonal physical map of the nematode genome is nearly complete<sup>18</sup>. A combination of cosmid and yeast artificial chromosomal (YAC) clones has been used to reconstruct more than 95 megabases (Mb) of the genome. More than 90 Mb have been positioned along the chromosomes using genetically mapped sequences and through *in situ* hybridization of cloned sequences. Fewer than 40 gaps now remain in the map and progress towards closure is proceeding steadily (A.C. *et al.*, unpublished results).

We have embarked on a project to determine the entire sequence of the nematode genome. The region where we began is the centre of chromosome III (Fig. 1). There is good cosmid coverage over most of several megabases, and the few areas lacking cosmid coverage are spanned by YACs. The region lies in the central gene-rich cluster of chromosome III, where several genes of interest have been mapped.

## Strategy and methods

The physical map of *C. elegans* was constructed using cosmid and YAC clones, but directed sequencing of even cosmid clones proved impractical because of the presence of repeat sequences and problems with obtaining sufficient quantities of template DNA. Thus we began by generating random subclones from



sheared, sized DNA<sup>23,24</sup>. Libraries of small inserts (1–3 kilobases (kb)) were convenient and larger insert libraries (6–9 kb) were useful for establishing continuity in gap closure.

To collect the sequence data, we relied largely on two fluorescent-based sequence-gel readers, the Applied Biosystems ABI 373A and the Pharmacia ALF, which provide data directly in machine-readable form. All data were transferred to Unix-based Sun workstations. A display editor was developed to allow rapid clipping of the vector from the 5' end and unreliable sequence from the 3' end<sup>25</sup>.

In the first phase of data generation, single reads of about 400 base pairs (bp) were taken from one end of the random clones using the ABI 373 instrument and *Taq* polymerase with a cycle sequencing protocol that reduced the amount of template necessary<sup>26,27</sup>. After 100–350 reads were obtained (the optimal number will ultimately depend on relative costs and has not been established; Table 1) and assembled into contigs using Staden's assembly program<sup>28</sup>, the project switched to a directed phase for closure and finishing. As a preliminary directed step, a reverse read was sometimes obtained from the opposite end of selected inserts to help establish linkage, double stranding and gap closure. Further directed sequencing required custom oligonucleotide primers, whose selection was aided by OSP (for oligonucleotide selection program)<sup>29</sup>. For technical reasons, the Pharmacia ALF was more convenient for reads from custom primers<sup>30</sup>. After gap closure and double stranding, in some cases further reads had to be taken, often with different chemistries<sup>31,32</sup>, to resolve ambiguities. Final editing, and indeed editing throughout the project, was assisted by the ability to recall the original trace data for any region of concern from the editor program. Further details of our sequencing strategy will be published elsewhere<sup>33,34</sup>.

Once the final sequence was obtained, the databases were searched for similarities using the algorithm BLAST<sup>35</sup>. In addition, we have started to interpret the sequence directly. The program GENEFINDER (P.G. and L.H., unpublished) uses a statistically rigorous treatment of likelihoods to find possible genes. Other features, such as repeated sequences, are also being examined. The annotated sequences have been submitted to Genbank and EMBL databases. To present the sequences in

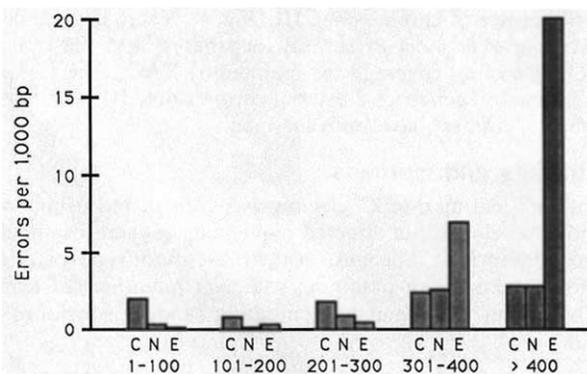


FIG. 2 The types of errors found when comparing the initial machine reads with the final edited sequence. The errors are broken down into the following categories: errors due to either the enzyme or the gel, resulting in compressions or stops (C) and errors due to the software, where there was either no call made (N, nocall) or an error made in base calling (E). In the last case the software has either inserted an extra base (an overcall), failed to recognize a base (an undercall) or simply called the wrong base (a miscall). Nocalls are less troublesome than an error in base calling, and nocalls accounted for most of the software failures in the first 300 bases of the reads. Software errors, particularly base calling errors, predominated above 300 bases. The majority of these errors were overcalls (87), as opposed to undercalls (20) or miscalls (16). The data are taken from 89 reads of single-stranded templates, with average length of 427 bases (Fig. 1), done on cosmid F59B2, for which changes in our data-handling software made the analysis easier. The data quality, however, should be similar in the cosmid sequences presented here.

TABLE 1 Sequencing strategies and statistics

	B0303	ZK637	ZK643
Cosmid clone	B0303	ZK637	ZK643
Insert DNA size (bp)	41,071	40,699	39,528
Random subclone libraries size range:	5–6 kb	9–14 kb	1–2 kb, 6–9 kb, 9–14 kb
cloning vector:	pUC118 <sup>47</sup>	pBS <sup>48</sup>	M13mp18 <sup>49</sup> , pEMBL9, pBS
Random sequencing method	ds, ABI	ds, ABI; ss, <sup>32p</sup>	ss, ABI
Number of random subclones	197	102	360
Number of reverse primer readings	119	70	0
Closure method	ss/ds, ALF	ss, <sup>32p</sup>	ss, ALF
Oligonucleotide primers required	102	417	100
Total number of readings	440	589	496
Average bp per read (vector sequence removed)	415	339	390
Total bp read (for assembly)	171,375	184,000	186,437
Final sequence redundancy	3.6	3.4	4.3

Abbreviations: ss, single-stranded template; ds, double-stranded template; ABI, Applied Biosystems Inc. 373A sequence-gel reader; ALF, Pharmacia ALF sequence-gel reader.

the context of our knowledge about the worm, we developed a *C. elegans* database, ACEDB, which holds not only the available sequences for the nematode, but also physical and genetic map information, along with reference lists and strain information (R.D. and J.T.-M., unpublished).

## Sequences

Three cosmids have so far been completed (sequences submitted to the EMBL and Genbank databases) during the development of our strategy, each with method variations to improve efficiency (Table 1). The first cosmid, ZK637, was sequenced using a minimum of random clones and radioactively labelled primers for walking. The number of primers required proved costly, and the use of films to collect data made editing difficult. ZK643 was partly sequenced using restriction enzyme partial digestion for the random clone production: although subclone recovery was high, these clones were biased in their representation. B0303 DNA was sheared to produce the subclones and, as for ZK637, reverse primer sequencing was used to establish linkage before beginning directed sequencing. Closure for both ZK643 and B0303 was done using directly labelled primers on the Pharmacia ALF.

The overlap of the manual, radiolabelled sequence with the fluorescent shotgun reads in ZK637 amply confirms the fidelity of the fluorescent method. The accuracy of the base calling was tested by comparing a sampling of individual sequence reads with the final edited sequence (Fig. 2). Generally, the only errors found in the first 300 bases of a read could be attributed to compressions or stops due to gel or enzyme limitations. Above 300, software limitations predominate, with increasing numbers of sites at which either no base call can be made or errors in the base calling occur. The failure to call a base is not a serious problem as no false information enters the database. By far the most common error is overcalling the number of bases in a run, but undercalls and miscalls also occur; these errors are easily corrected, once attention has been drawn to them, by reference to the traces.

The accuracy of the final sequence is difficult to estimate without extensive independent tests, but several factors give us confidence that the fidelity to the true genomic sequence is high. All sequences were determined at least once on both strands, with the exception of certain long tandem repeat sequences. Regions with compressions or other unresolved conflicts were resequenced. Possible errors indicated by the similarity searches

and GENEFINDER analysis were checked. To detect major cloning artefacts, we compared the polymerase chain reaction products derived from the cosmids with both the predicted lengths and the products from genomic DNA across the sequenced regions. The walking primers often proved useful for this purpose. For one cosmid, part of the vector sequence was analysed which was derived from the random phase alone so that not every region had been sequenced on both strands; here only one base in 2,000 was at variance with the available sequence.

Using similarity searches and GENEFINDER to analyse these sequences, a high density of likely genes was revealed (Fig. 3). One of the most striking similarities was found between the 22–26-kb region of ZK637 and the 116K subunit (relative molecular mass 116,000) of the rat vacuolar proton pump (Fig. 4). The similarity falls into several blocks which GENEFINDER suggests are probably exons, and spans from residues 14 to 824 of the 838-amino-acid protein. The amino-terminal half is most highly conserved between the two sequences; in one stretch 99 of 139 residues (71%) are identical and in a second, 223 of 314 (71%) are identical, with no gaps introduced into the alignment. The carboxy-terminal half has less similarity but contains sequences similar to all eight postulated transmembrane domains. Other extensive similarities were found to acetyl-CoA acetyltransferase, glutathione reductase, the arsenical pump-driving ATPase, the hypothetical transposase TcA of the nematode transposon Tc1, and the host protective factor of the parasitic nematode *Trichostrongylus colubriformis*. Less extensive but still highly significant similarities (BLAST scores of >100 over one or two exons) were found with the 50S ribosomal protein L11 and the neutrophil oxidase factor. This last similarity includes a motif shared between the oxidase factor, yeast actin-binding protein ABP-1, acanthamoeba myosin IC and *sre*-related kinase<sup>36</sup>. The hypothetical nematode protein has three copies of this motif. Finally, other more limited but still significant similarities of likely *C. elegans* coding regions were found with phenylethanolamine-*N*-methyltransferase, adenylyl cyclase, giant secretory protein, the yeast *CDC25* cell-cycle gene (and the *Drosophila* string homologue), glucose transporter and immediate early protein IE110 of herpes simplex virus. For adenylyl cyclase, several adjacent segments of B0303 showed similarity with a repeated motif of the cyclase sequence.

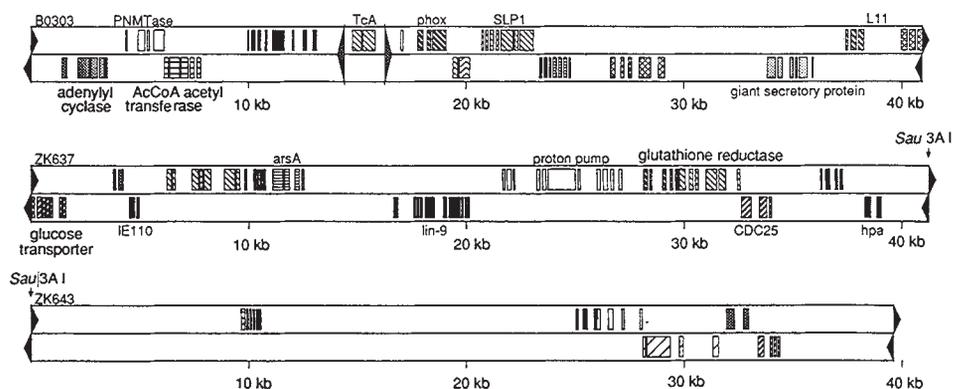
Altogether, GENEFINDER predicts a total of 33,573 bases (27%) of coding sequence in the total of 121,298 bases sequenced. For the individual cosmids the percentage of coding sequences are 37% for B0303, 33% for ZK637, and 13% for ZK643. The number of different genes represented is difficult to estimate,

as at present there are no clear rules by which to distinguish ends of genes from introns. Taking into account factors such as the distinct homologies and the spacing and strandedness of exons, we would estimate B0303 contains 14 genes, ZK637 12 genes and ZK643 6 genes. Transcript analysis will be required to test these predictions. Several studies have already centred on genes in the region. The gene *sup-5* (ref. 37) had been mapped close to the end of B0303; indeed, the B0303 sequence overlaps with the 4.2-kb region previously sequenced around the transfer RNA gene *sup-5* (K. Kondo and R.W., unpublished results). The genes *lin-9* (ref. 38) and *unc-32* (ref. 6) were known by transformation rescue to lie within ZK637 before sequencing began. The gene *lin-9* has been shown to correspond to the predicted gene on the bottom strand of ZK637 in the region 20.2–16.6, and the complementary DNA sequence used to refine the predictions of the GENEFINDER analysis (G. Beitel and R. Horvitz, personal communication). From its genetic position, *unc-32* is likely to be one of the genes immediately to the right of *lin-9*.

The sequences have also been scanned for repeats. Large, almost perfect inverted repeats (465/468) flank the region of B0303 near the 15-kb point (Fig. 3) that shares similarity with the TcA transposase; subsequent analysis revealed this region to represent a copy of the Tc1-related transposon Tc3 (D. Schneider, J. Collins and P. Anderson, personal communication). Examples of short tandemly repeated sequences include 21 copies of a 59-bp motif at 35 kb in ZK643 and two segments (11 and 13 copies) of an 11-bp motif inverted with respect to one another at 13.3 and 13.7 kb in ZK637. A larger duplicated segment is present spanning the join of ZK637 and ZK643, where blocks of 99, 305 and 789 bp spread over 1.6 kb are almost exactly duplicated 4 kb away. A 500-bp region near 7 kb in ZK643 shares several fragments with strong homology (>90%) to an 800-bp region near 32 kb in B0303, the order and orientation of the fragments being different in the two cases. A 1-kb region near 38-kb in ZK643 contains large stretches of the nucleotide motif NGG tandemly repeated; the same motif is found at the fragile X site<sup>39–41</sup>. Some 20 copies of a 94-bp consensus sequence were found dispersed throughout the three cosmids, many in inverted pairs separated by up to 130 bp. Some copies showed a good match to the consensus (80–90% identity), whereas others diverged strongly or were incomplete. A search of Genbank showed the sequence to be specific to *C. elegans* and present in four copies in other entries. Although they are relatively few compared with those in mammalian DNA and are confined to local regions, these repeats would make impossible a walking strategy that relied primarily on cosmid templates.

FIG. 3 The genes in the cosmids sequenced, as determined through homology searches, GENEFINDER analysis, and available cDNA sequences. The exons of predicted genes are indicated by shaded blocks, with different shading patterns indicating distinct genes. Those shown above centre for each cosmid are encoded by the top strand and those below by the bottom strand. The genes with significant similarities to genes in the databases are indicated by the name of the most similar sequence (see Fig. 4 for details). Also shown is the position of the *lin-9* gene, as determined from cDNA sequence (G. Beitel and R.

Horvitz, personal communication). ZK637 and ZK643 overlap by the single *Sau3A* site indicated. The inverted repeats flanking the TcA homologue are indicated by shaded triangles. The region at 25.1 to 28.1 kb of ZK643 contains two predicted genes that overlap by 52 bases but use distinct



reading frames in the overlapped segment. This is indicated in the diagram by different shading within the same block, but has not been confirmed by experiment. PNMTase, phenylethanolamine-*N*-methyltransferase; phox: neutrophil oxidase factor.



11. Sulston, J. E., Schierenberg, E., White, J. G. & Thomson, J. N. *Devil Biol.* **100**, 64–119 (1983).
12. White, J. G., Southgate, E., Thomson, J. N. & Brenner, S. *Phil. Trans. R. Soc.* **314**, 1–340 (1986).
13. Emmons, S. W., Yesner, L., Ruan, K. S. & Katzenberg, D. *Cell* **32**, 55–65 (1983).
14. Eide, D. J. & Anderson, P. *Proc. natn. Acad. Sci. U.S.A.* **82**, 1756–1760 (1985).
15. Moerman, D. G., Benian, G. M. & Waterston, R. H. *Proc. natn. Acad. Sci. U.S.A.* **83**, 2579–2583 (1986).
16. Coulson, A. R., Sulston, J. E., Brenner, S. & Karn, J. *Proc. natn. Acad. Sci. U.S.A.* **83**, 7821–7825 (1986).
17. Coulson, A. R., Waterston, R. H., Kiff, J. E., Sulston, J. E. & Kohara, Y. *Nature* **335**, 184–186 (1988).
18. Coulson, A. *et al. BioEssays* **13**, 413–417 (1991).
19. Stinchcomb, D. T., Shaw, J. E., Carr, S. H. & Hirsh, D. I. *Molec. cell. Biol.* **5**, 3484–3496 (1985).
20. Fire, A. *EMBO J* **5**, 2673–2680 (1986).
21. Fire, A. & Waterston, R. H. *EMBO J* **8**, 3419–3428 (1989).
22. Blumenthal, T. & Thomas, J. H. *Trends Genet.* **4**, 305–308 (1988).
23. Schriefer, L. A., Gebauer, B. K., Qiu, L. Q. Q., Waterston, R. H. & Wilson, R. K. *Nucleic Acids Res.* **18**, 7455–7456 (1990).
24. Deininger, P. L. *Analyt. Biochem.* **129**, 216–223 (1983).
25. Gleeson, T. & Hillier, L. *Nucleic Acids Res.* **19**, 6481–6483 (1991).
26. Smith, L. M. *et al. Nature* **321**, 674–679 (1986).
27. Craxton, M. *Methods: A Companion to Methods in Enzymology* **3**, 20–26 (1991).
28. Dear, S. & Staden, R. *Nucleic Acids Res.* **19**, 3907–3911 (1991).
29. Hillier, L. & Green, P. *PCR Meth. Appls* **1**, 124–128 (1991).
30. Ansonge, W., Sproat, B. S., Stegemann, J. & Schwager, C. *J. biochem. biophys. Meth.* **13**, 315–323 (1986).
31. Mizusawa, S., Nishimura, S. & Seela, F. *Nucleic Acids Res.* **14**, 1319–1324 (1986).
32. Hawkins, T. L. & Sulston, J. E. *Nucleic Acids Res.* **19**, 2784 (1991).
33. Hawkins, T. L., Du, Z., Halloran, N. D. & Wilson, R. K. *Electrophoresis* (manuscript submitted).
34. Craxton, M. in *DNA Sequencing: Laboratory Protocols* (eds Griffin, H. G. & Griffin, A. M.) (Humana, NJ, 1992).
35. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. *J. molec. Biol.* **215**, 403–410 (1990).
36. Drubin, D. G., Mulholland, J., Zhu, Z. & Botstein, D. *Nature* **343**, 288–290 (1990).
37. Willis, N. *et al. Cell* **33**, 575–583 (1983).
38. Ferguson, E. L. & Horvitz, H. R. *Genetics* **110**, 17–72 (1985).
39. Oberle, I. *et al. Science* **252**, 1097–1102 (1991).
40. Yu, S. *et al. Science* **252**, 1179–1181 (1991).
41. Verkerk, A. J. M. H. *et al. Cell* **65**, 905–914 (1991).
42. Herman, R. K. in *The Nematode Caenorhabditis elegans* (eds Wood, W. B. *et al.*) 17–45 (Cold Spring Harbor Laboratory, New York, 1988).
43. Heine, U. & Blumenthal, T. *J. molec. Biol.* **188**, 301–312 (1986).
44. Olson, M. in *Genome Dynamics, Protein Synthesis and Energetics* (eds Broach, J. R., Pringle, J. R. & Jones, E. W.) 1–41 (Cold Spring Harbor, NY, 1991).
45. Hall, L. M. C., Mason, P. J. & Spierer, P. *J. molec. Biol.* **169**, 83–96 (1983).
46. Bossy, B., Hall, L. M. C. & Spierer, P. *EMBO J* **3**, 2537–2541 (1984).
47. Vieira, J. & Messing, J. *Meth. Enzym.* **153**, 3–11 (1987).
48. Short, J. M., Fernandez, J. M., Sorge, J. A. & Huse, W. D. *Nucleic Acids Res.* **16**, 7583–7600 (1988).
49. Yanisch-Perron, C., Vieira, J. & Messing, J. *Gene* **33**, 103–119 (1985).
50. Burglin, T. R., Finney, M., Coulson, A. & Ruvkin, G. *Nature* **341**, 239–243 (1989).
51. Pearson, W. R. & Lipman, D. J. *Proc. natn. Acad. Sci. U.S.A.* **85**, 2444–2448 (1988).
52. Kataoka, T., Broek, D. & Wigler, M. *Cell* **43**, 493–505 (1985).
53. Kaneda, N. *et al. J. Biol. Chem.* **263**, 7672–7677 (1988).
54. Sasaoka, T., Kaneda, N., Kurosawa, Y., Fujita, K. & Nagatsu, T. *Neurochem. Int.* **15**, 555–565 (1989).
55. Fukao, T. *et al. J. Biochem., Tokyo* **106**, 197–204 (1989).
56. Prasad, S. S., Harris, L. J., Baillie, D. L. & Rose, A. M. *Genome* **34**, 6–12 (1991).
57. Leto, T. L. *et al. Science* **248**, 727–730 (1990).
58. Wada, Y., Kitamoto, K., Kanbe, T., Tanaka, K. & Anraku, Y. *Molec. cell. Biol.* **10**, 2214–2223 (1990).
59. Lendahl, U. & Wieslander, L. *Cell* **36**, 1027–1034 (1984).
60. Dognin, M. J. & Wittman-Liebold, B. *Eur. J. Biochem.* **112**, 131–151 (1980).
61. Post, L. E., Strycharz, G. D., Nomura, M., Lewis, H. & Dennis, P. P. *Proc. natn. Acad. Sci. U.S.A.* **76**, 1697–1701 (1979).
62. Downing, W. L., Sullivan, S. L., Gottesman, M. E. & Dennis, P. P. *J. Bact.* **172**, 1621–1627 (1990).
63. Sauer, N., Friedl, K. & Wicke, U. Genbank Accession Number X55350 (1991).
64. McGeoch, D. J. *et al. J. gen. Virol.* **69**, 1531–1574 (1988).
65. Perry, L. J., Rixon, F. J., Everett, R. D., Frame, M. C. & McGeoch, D. J. *J. gen. Virol.* **67**, 2365–2380 (1986).
66. Chen, C.-M., Misra, T. K., Silver, S. & Rosen, B. P. *J. Biol. Chem.* **261**, 15030–15038 (1986).
67. Perin, M. S., Fried, V. A., Stone, D. K., Xie, X.-S. & Sudhof, T. C. *J. Biol. Chem.* **266**, 3877–3881 (1991).
68. Greer, S. & Perham, R. N. *Biochemistry* **25**, 2736–2742 (1986).
69. Russell, P. & Nurse, P. *Cell* **45**, 145–153 (1986).
70. Frenkel, M. J., Dopheide, T. A., Wagland, B. M. & Ward, C. W. Genbank Accession Number M63263 (1991).

ACKNOWLEDGEMENTS. We thank G. Beitel and R. Horvitz for information on *lin-9*; P. Anderson for sequence of Tc3; M. Jier and R. Showkneen for synthesis of oligonucleotides; and P. Kassos for preparing the manuscript. The work was supported by grants from the NIH Human Genome Center and the MRC HGMP, as well as our respective institutions.

## LETTERS TO NATURE

## Vortices on accretion disks

M. A. Abramowicz, A. Lanza, E. A. Spiegel\*† & E. Szuszkiewicz‡

Scuola Internazionale Superiore di Studi Avanzati, Via Beirut 4, 34014 Trieste, Italy

NORDITA, Blegdamsvej 17, 2100 Copenhagen, Denmark

ICTP, Strada Costiere, 11, 34014 Trieste, Italy

\* Columbia University, 538 West 120 Street, New York, New York 10027, USA

† University of Ferrara, Via Paradiso, 12, 44100 Ferrara, Italy

‡ Queen Mary and Westfield College, Mile End Road, London E1 4NS, UK

**EVERY rotating cosmic fluid that can be observed sufficiently closely displays either vortices or magnetic flux tubes on its surface; examples are tornadoes in the Earth's atmosphere<sup>1</sup>, the Great Red Spot and other vortices in Jupiter's atmosphere, and sunspots. We suggest here that hot accretion disks also produce coherent objects, and that these vortices and magnetic flux tubes will cause significant dissipation and other observable physical effects. They will facilitate the escape of collimated radiation from deep within hot disks, producing spectral changes and time variability in the radiation from the disk. In the case of active galactic nuclei, modification of X-ray spectra due to the presence of vortices on accretion disks permits us to explain several observational puzzles, including short-term variability and the low degree of linear polarization.**

Coherent structures are known to be common on planets and stars. As well as the examples first given, it is generally accepted that structures related to sunspots occur on other cool stars<sup>3</sup>, and there is growing evidence for spots on hot stars<sup>4</sup>. Laboratory experiments intended to simulate planetary atmospheres produce long-lived coherent objects under suitable conditions<sup>5,6</sup>. There seems to be no natural occurrence of rotating turbulence without ordered structures embedded in it. On this basis, we must expect that accretion disks are dotted with an admixture of vortices and magnetic flux tubes.

The direct consequences of vortices and magnetic flux tubes

are most impressive when there are secondary constituents in the material of the disks. In the primitive solar nebula, the formation of dust<sup>7</sup> is influenced by strong vortices in whose cores the temperature may be significantly lowered<sup>8</sup>. The dust can clump between the vortices on being extruded from them by centrifugal force. Likewise, in the case of hot disks, like those in active galactic nuclei (AGN), the emergent radiation will be strongly affected.

Both general considerations<sup>9</sup> and elementary transfer theory<sup>10</sup> show that vortices and magnetic flux tubes will concentrate emerging radiation into strong beams. Photons will tend to seek the low-pressure regions inside the tubes, such as hydrogen bubbles introduced into a turbulent fluid in laboratory experiments migrate into the cores of strong vortices<sup>11</sup>. Once in the low-density region, the photons quickly make their way out of the disks directly from their relatively hot interiors. This radiation will be collimated and considerably harder than the radiation from a disk with no vortices, and it can provide a substantial fraction of the total luminosity of the disk. The outflowing radiation may be scattered on cooler external material before reaching the observer. From these phenomenological considerations, we can understand a number of the observed properties of AGNs.

(1) The short-term X-ray variability of AGNs is featureless in the frequency range from  $10^{-3}$  Hz to  $10^{-5}$  Hz. Their smooth variability spectra are described by a power-law distribution. A suggested explanation<sup>12,13</sup> is that there are many small, well-separated bright spots on the disk surfaces. When, on a relativistically rotating accretion disk seen at a non-zero viewing angle, an individual spot periodically moves towards the observer, its brightness is Doppler-amplified. This produces a single line (or lines) in the variability power spectrum corresponding to the orbital frequency (and its higher harmonics). Spots distributed at a range of radii produce many lines which overlap to form a power-law spectrum<sup>12,13</sup>. This empirical picture is physically natural if strong vortices or flux tubes are present.

(2) The linear polarization of most of the quiescent quasars is  $\sim 1\%$  or lower. In the standard thin disk model, in which