Minireview

Yeasties and beasties: 7 years of genome sequencing

Karen Thomas

The Sanger Centre, Wellcome Trust Genome Campus, Hinxton CB10 1SA, UK

Received 10 September 1996

Abstract The *Saccharomyces cerevisiae* genome sequencing project was the first of many projects aimed at sequencing the entire genomes of model organisms. Since its initiation in 1989, there have been numerous debates about the validity of genome sequencing, especially with reference to the model organisms. Seven years on, I hope to satisfy some of the critics by demonstrating that, as a consequence of the mass of data now becoming available from such projects, and the beginning of the major collaborative effort to sequence the human genome, we are now entering an exciting and dynamic time for those involved not only in genome sequencing, but also in all areas of the biological sciences.

Key words: Model organism; Genome sequencing; Shotgun sequencing

1. Introduction

The concept of sequencing the genomes of model organisms such as *Saccharomyces cerevisiae*, (12.5 Mb), *Escherichia coli* (4.7 Mb) and *Caenorhabditis elegans* (100 Mb) arose at a time when tackling the entire human genome seemed beyond the scope of existing technology. Smaller, less complex genomes, offered the opportunity of developing the technologies required for sequencing the human genome whilst generating fascinating biological information. The decision to set aside large amounts of research funding to these projects was controversial. Seen as diverting resources from basic biological research, it could only be justified if biological research as a whole were seen to benefit.

2. Historical perspectives

In 1989, Andre Goffeau set up an EU consortium to sequence the genome of the budding yeast *S. cerevisiae* [1]. Many of the 74 laboratories involved were not yeast laboratories, but were drawn to the project in the hope of sequencing the homologs of their favourite genes. Most of the laboratories used a shotgun sequencing approach (Fig. 1) involving two stages; an initial 'random' stage of data collection, followed by a directed primer walking stage to complete the sequence. The initial random stage can involve the generation of clones from whole genomes, or fragments of variable size (from small restriction fragments to entire chromosomes). In the case of the yeast, most laboratories worked from Goffeau's map of cosmid and lambda clones [2]. During the course of the project shotguns of whole chromosomes were attempted [3]. This approach met with problems due to difficulties in obtaining sufficiently pure individual chromosomes and was modified to include existing mapped clones where available.

1991 saw the start of four projects aimed at laying the foundations for human genome sequencing [4]. Three were USA-based projects, (*E. coli*, *S. cerevisiae* and *Mycoplasma capricolum*) the fourth (the nematode) was a joint collaboration between the Cambridge based laboratory of John Sulston and Bob Waterston's laboratory at the Washington University Medical School, St Louis. Table 1 outlines these and other model organism projects discussed in this review, and the strategies they employ. These include 'shotgun' sequencing (used by Fred Sanger to sequence the 48.5 kb bacteriophage λ genome in 1982 [5]) using a range of sources of DNA and multiplex walking using genomic DNA [6,7]. The *E. coli* pilot project worked from a physical map of overlapping lambda clones [8]. During initial technical developments, nine clones (100 kb) were sequenced by subcloning into m13, followed by radioactive sequencing. About 75% of the *E. coli* genome has now been completed and is accessible through public databases (Peter Sterk, personal communication).

Botstein and Davis also applied shotgun sequencing (using fluorescence sequencing machines) to part of the *S. cerevisiae* genome, completing some 1.1 Mb (8.4%) of the genome by 1995 [1,4]. Gilbert's group used the multiplex walking approach for the genome of *M. capricolum* [6,7]. Direct genome sequencing was used to circumvent the effort required in the preparation of a physical map and problems of non-random genome coverage thought to be associated with the process (due to repeats or 'poisonous' genes held within the clones [9]). They completed 250 000bp by 1995, an impressive total for a small laboratory, demonstrating that the multiplex approach is applicable to sequencing large chunks of genomic DNA. Although direct genome sequencing has been applied successfully to a number of simple genomes [10–12] there is a worry that the repeats associated with more complex genomes may prove intractable to such an approach.

Each of these projects demonstrated proof of principle in terms of technical ability. What they have failed to deliver is the scaling-up required to tackle a project the size of the human genome. In this respect they have been largely overshadowed by the evolution of a handful of specialised sequencing laboratories.

Not least of these, J. Craig Venter's laboratory at the Institute of Genome Research (TIGR), completed the first bacterial genomes; *Haemophilus influenzae* and *M. genitalium* for which whole genome shotgun sequencing was employed [10–12]. As can be seen from Table 1, the majority of current

PII S0014-5793(96)01063-0

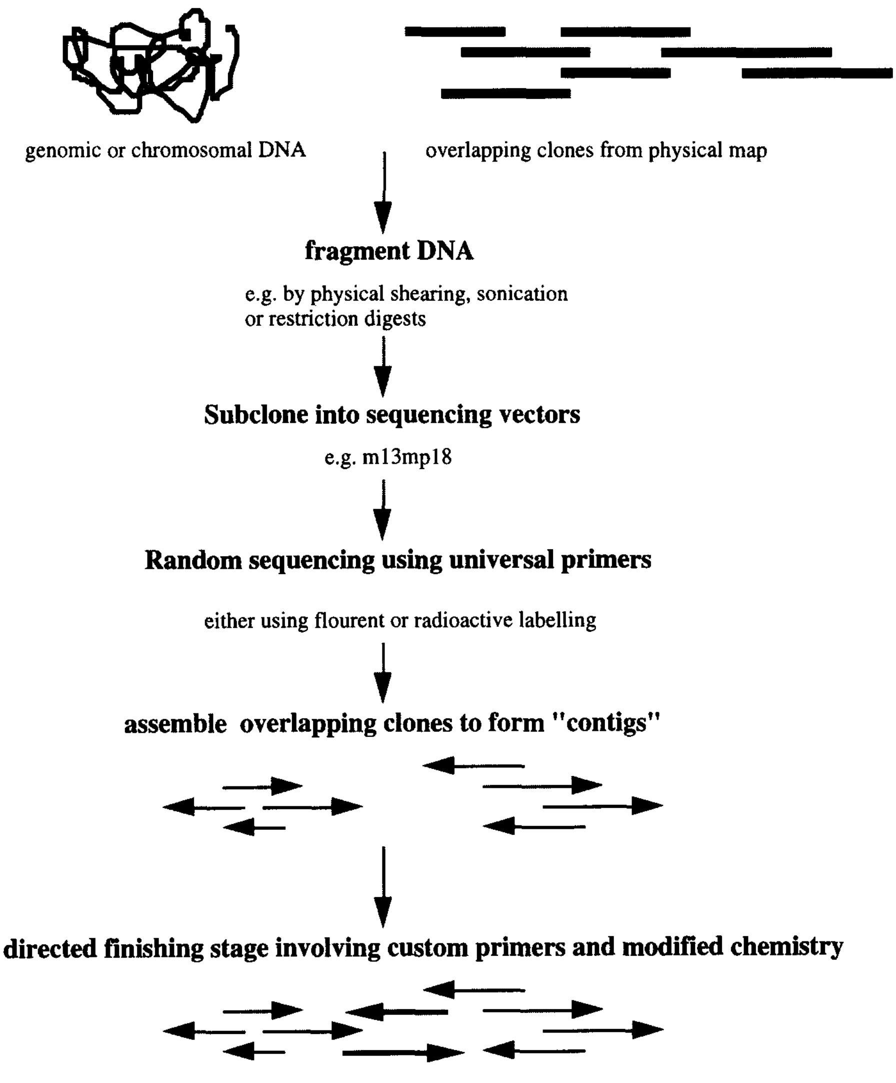
genomic or chromosomal DNA
overlapping clones from physical map
fragment DNA
e.g. by physical shearing, sonication
or restriction digests
Subclone into sequencing vectors
e.g. m13mp18
Random sequencing using universal primers
either using flourent or radioactive labelling
assemble overlapping clones to form "contigs"
directed finishing stage involving custom primers and modified chemistry

Fig. 1. Stages involved in the 'shotgun' sequencing approach.

genome projects use shotgun methodology, including that of *C. elegans*.

The roots of the *C. elegans* project lie in the collaboration between the laboratories of Bob Waterston and John Sulston, first to produce the nematode physical map, and then to sequence the genome. The map, which has provided the foundation of the sequencing project was started by John Sulston and Alan Coulson (the map's 'curator') in the mid 1980s. Sequencing began in 1991, with a 3 year pilot project to complete 3 Mb (3%) of the genome. The project was jointly funded by the NIH and MRC. One of the first areas to be tackled was sequence throughput.

The aim from the beginning was to scale-up to enable the completion of the entire 100 Mb of the nematode genome cheaply and efficiently. This has been achieved with very little change in basic approach. Improvements made to throughput and efficiency have resulted in the reduction of costs from $5 per base at the beginning, to the present cost of 40 cents per finished base. Over the years, all stages have become routine, and can be carried out by technical staff trained on site. As the experience of our teams has increased, so has output and efficiency. All this has been achieved (largely) without any significant level of automation, something that will provide further improvements in the future.

These developments were achieved in the first instance through the nematode genome, and subsequently expanded to include yeast (*S. cerevisiae* and *S. pombe*), human, and more recently pathogen genomes such as *Mycobacterium tuberculosis* and *Plasmodium falciparum*. Each genome has presented a different set of challenges in terms of both sequencing chemistry (e.g. GC-rich regions of the human CpG islands, and AT-rich sequence of the *P. falciparum* genome), or software challenges (e.g. handling alu repeats in human).

In 1992 the Wellcome trust committed to sequence up to 50% of the yeast genome, involving Bart Barrell's group (Sanger Centre) in a substantial expansion of effort. This was closely followed by a similar commitment by the NIH, which enabled Mark Johnston's group (St Louis) to join the project [13], working from Maynard Olsen's physical map [14,15]. The involvement of these two laboratories was just the push that the project needed; the yeast genome was completed early this year with substantial contributions from the EU consortium (55%), Sanger Centre (17%), and St Louis (10.4%) [1,16].

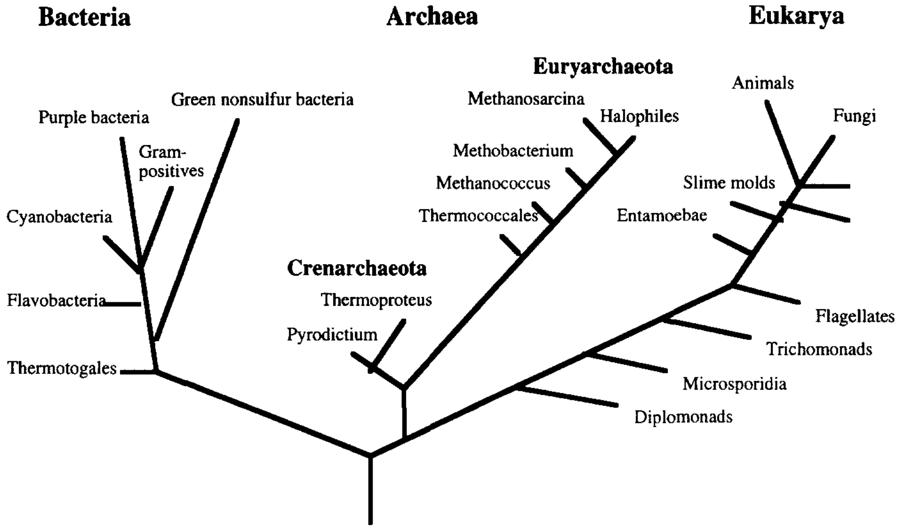

Fig. 2. Phylogenetic tree in rooted form, showing 3 domains. Branch lengths are based on upon rRNA sequence comparisons.

3. The impact of genome sequencing on biologists

The technological improvements fuelled by the sequencing projects are already finding their way into other laboratories. In addition, there has been a parallel advance in informatics, with a view to dealing with the amounts of data generated by such projects. Although I do not have space to cover this aspect in detail here, it must be said that this field has had a tremendous impact on the ability of the scientist in the field to both access and use the data available. Apart from this, how has all this data helped in our understanding of biology as a whole? A few examples of the impact of genome sequencing on the scientific community, are given below.

4. Genome sequencing and phylogeny

Recent weeks have seen the publication by TIGR of the genome sequence of the first archaeon, *Methanococcus jannaschii* [17], presenting scientists with a unique opportunity to compare genomes of organisms from the 3 major branches of life: archaea, bacteria, and eukarya [17,18].

In the late 1970s, Woese used comparison of tRNA sequences as a means of characterising a number of different microbial species [19]. In doing so he discovered a new group of organisms which were morphologically prokaryotes, but had the RNA profile of eukaryotes [19,20]. This data effectively split the prokaryotes into the true bacteria (or eubacteria), and a second group (the archaebacteria) containing the methanogenic bacteria, the thermoacidophiles, and the extreme halophiles. This went against the accepted view of division into prokaryotes (organisms with no nuclear membrane), and eukaryotes (organisms with a distinct nuclear membrane) [21].

The complete sequence of the *M. jannaschii* genome (1.7 Mb), yielded some 1734 predicted genes of which 56% had no known database similarities. As predicted by Woese, those genes with database matches were much more closely related to eukaryotic genes than prokaryotic genes, establishing once and for all the existence of the third phylogenetic domain (Fig. 2).

5. Inter-specific sequence comparisons and human inherited diseases

Use of sequence homology between distantly related species (e.g. hybridisation to zoo-blots) has been used for some time

Table 1
Genome sequencing projects in this review

Organism	Genome size (Mb)	Description	Sequencing approach
M. jannaschii	1.7	archaea	whole genome shotgun
M. genitalium	0.6	prokaryote	whole genome shotgun
M. capricolum	0.8	prokaryote	multiplex walking
H influenzae	1.8	prokaryote	whole genome shotgun
M tuberculosis	4.4	prokaryote	map, shotgun
E. coli	4.7	prokaryote	map, shotgun
P falciparum	27	protozoan	whole chromosome+map, shotgun
S cerevisiae	12.5	single-celled eukaryote	map, shotgun
S pombe	14	single-celled eukaryote	map, shotgun
C. elegans	100	small metazoan	map, shotgun
D. melanogaster	165	small metazoan	map, directed sequencing
H. sapiens	3000	large metazoan	various

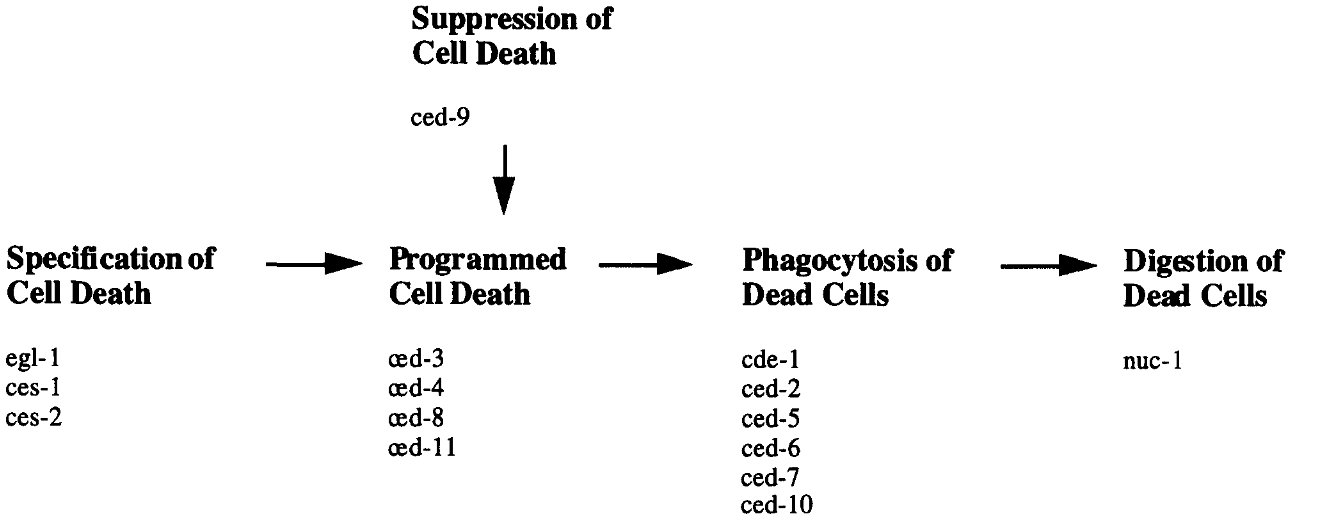

Fig. 3. Programmed cell death in *Caenorhabdis elegans*. The genes involved in the process of cell death determine whether or not a particular cell or class of cells die.

to aid the identification of coding sequences. Today, thanks to the wealth of genomic and cDNA data, this is possible at the sequence level.

Drosophila melanogaster is a unique model organism, due to the high number of well-characterised mutant genes. Information about more than 9000 genes and 25 000 alleles is held within 'Flybase', which contains information on all aspects of *Drosophila* genetics and molecular biology [22,23]. Despite the evolutionary distance between the human and *Drosophila*, these 2 metazoan genomes show remarkable conservation between genes. Indeed, there are already examples of mutations in homologous genes showing similar phenotypes in both organisms, as illustrated by the *haywire*, the fly homologue of ERCC3, a human excision repair gene that is implicated in Codayne's syndrome and xeroderma pigmentosum [24]

As a result of the Merck/Washington University expressed sequence tag (EST) project there has been a remarkable increase in the number of human ESTs deposited in the public database dbEST [25]. The 343 000 entries now include some 30% previously held in TIGRs private databases. Banfi et al. [26,27], exploited this remarkable resource, in a systematic search for potential human homologues of *Drosophila* genes with known mutant phenotypes. They screened dbEST entries from the Merck/Washington University consortium for matches to *Drosophila* mutant genes [28]. This approach produced some 66 human ESTs, which they called DRES clones (*Drosophila*-Related Express Sequences) which were mapped back to the human genome using FISH and mapping [29]. The On-line Mendelian Inheritance in Man (OMIM) db was used to retrieve information about inherited disease loci mapped to the same chromosome in an attempt to identify DRES clones that provided positional candidate genes for human diseases.

One example, DRES9, homologous to *Drosophila* retinal degeneration B gene, mapped to 11q13.5, a region to which at least 3 types of human retinopathies are assigned. Radiation mapping data for this clone revealed a link with the marker D11S913, which showed strong genetic linkage with the locus of a retinopathy, Bardet-Biedl syndrome. Another example, DRES12 (homologous to the *Drosophila* eyes absent gene), mapped to human chrom 20q13.1, syntenic with mouse chromosome 2;G-H1. Screening of the Dysmorphic Human-mouse Homology Database (DHMHD) at the institute of Child Health (University of London) revealed that the mouse mutant blind-sterile mapped to the same chromosomal region [30].

The group are now working on the isolation of murine homologues to DRESs in order to perform genetic mapping and detailed expression studies in the mouse (Banfi and Borsani, personal communication); an elegant demonstration of how the available information from a number of different organisms can be brought together to provide a greater understanding of human genetic diseases.

6. Programmed cell death in *Caenorhabditis elegans*

Programmed cell death (apoptosis) plays an important role in development of multi-celled organisms. The nematode *C. elegans* is a key to our understanding of this process.

The nematode, a simple multicellular organism, is a very powerful model system for the study of the cell and molecular biology of multicellular animals. It has a well studied development from single-cell to adult, greatly enhanced by its transparency, which allows every stage of development to be studied in the live animal under Normarski light microscopy. Every cell movement, division and death during its development has been determined. During nematode development, 131 of the 1090 somatic cells of the nematode undergo programmed cell death [31,32]. Mutant nematodes with defects in these pathways have been identified, along with the gene(s) responsible, and this has led to a greater understanding of the process of apoptosis.

The genes involved in this process can be divided into 5 main groups (Fig. 3), responsible for triggering cell death, for the process itself, for engulfment, and finally for the disposal of the dead cell. Several of the genes involved in the pathway have been cloned and sequenced, either directly, or as a result of the genome sequencing project, and comparisons with homologous genes in other animals are leading to a better understanding of cell death in mammals. ced-9, is required to prevent activation of apoptosis in *C. elegans*. Mutations that cause gain-of function, or over-expression of the wild-type ced-9 genes result in the survival of cells that normally die. Mutations reducing ced-9 expression or function, causes the death of many cells that normally live, leading to embryonic lethality [33]. Clearly, the activities of ced-9 and other cell death genes are essential for normal nematode development.

Similar genes have been identified in vertebrates and viruses, including bcl-2 a proto-oncogene which demonstrates striking functional similarity to ced-9. Over expression of bcl-2 protects from cell death, and reduction/loss of function causes cells to be hypersensitive to cell death-inducing signals. Most striking is the observation that human bcl-2 can prevent programmed cell death in *C. elegans* [34]. Sequence comparison between the *C. elegans* ced-9 and mouse bcl-2 shows an overall identity of 24% (49% similarity). bcl-2 belongs to a rapidly growing family of genes with similarity to other ced-9 and other nematode cell death genes. ced-3 is homologous to the interleukin-1β converting enzyme (ICE) family of cysteine proteases in mammals. As yet no known homologues have been found for ced-4, which contains potential calcium binding domains, but shows no significant sequence similarity to known mammalian proteins; it may be related to an as yet unknown family of mammalian proteins [35]

The *C. elegans* apoptosis pathway promises to be particularly useful in highlighting the basic elements of the complex interacting pathways involved in mammalian apoptosis and the identification of the genes involved.

5. Summary

It is 7 years now since the beginning of the first model organism genome sequencing project. Those 7 years have seen the completion of the first bacterial, eukaryotic and archaeal genomes. The first metazoan project (the nematode) is now half-way to completion, and will be complete in 2 years time. We have seen an improvement in the cost and efficiency of genome sequencing to the level at which we can consider tackling the complete human genome.

Data provided by the model organisms has greatly enhanced our understanding of many different biological systems, and promises to improve our ability to identify human disease genes. The examples above are wonderful demonstrations of how all the resources, be they computational, libraries, maps or models are now creating a powerful battery of techniques to investigate human genetic diseases and biological processes alike

Far from exhausting their potential, the model organisms becoming increasingly important, and the field is continuing to expand, incorporating organisms as diverse as pathogens, plants (*Arabidopsis* and mice. As more human genomic DNA sequence becomes available, it is becoming clear that both genomic and est data from all the model organisms, are of great help in predicting coding sequences. Furthermore, sequence data from vertebrate genomes such as the puffer fish (*Fugu rubripes*, [36]) and ultimately the mouse will be invaluable.

Since 1991, mapping of the human genome has moved on apace, and in May 1995 at the annual Genome mapping and sequencing project at Cold Spring Harbor, the decision was made to embark on the complete sequencing of the human genome. To this end a large collaborative network has been set up with the division of the genome between the established large-scale sequencing laboratories. We are entering an exciting new era, and must conclude that model organism sequencing has more that fulfilled its early promise. Hopefully, those early critics will now agree.

References

[1] Goffeau, A. (1994) Nature 369, 101–102.

[2] Thierry, A., Gaillon, L., Galibert, F. and Dujon, B. (1995) Yeast 11, 121–135.

[3] Churcher, C., Bowman, S., Badcock, K., Bankier, A., Brown, D., Connor, R., Devlin, K., Gentles, S., Hamlyn, N., Harris, D., Horsnell, T., Hunt, S., Jagels, S., Jones, M., Lye, G., Moule, S., Moule, T., Odell, C., Pearson, D., Rajandream, M., Rice, P., Rowley, N., Skelton, J., Smith, V., Walsh, S., Whitehead, S. and Barrell, B. (1996) Nature, Submitted.

[4] Roberts, L. (1991) Science 250, 1336–1338.

[5] Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F. and Peteson, G.B. (1982) J. Mol. Biol. 162, 729–773.

[6] Dolan, M., Ally, A., Purzycki, M.S., Gilbert, W. and Gillevet, P.M (1995) BioTechniques 19, 264–273.

[7] Ohara, O., Dorit, R.L. and Gilbert, W. (1989) Proc. Natl. Acad. Sci. USA 86, 6883–6887.

[8] Daniels, D.L., Plunkett, G., Burland, V. and Blattner, F. (1992) Science 257, 771–778.

[9] Szybalski, W. (1993) Gene 135, 279–230.

[10] Fraser, C.M., et al. (1995) Science 270, 397–403.

[11] Fleischman, R.D. et al. (1995) Science 269, 496–512.

[12] Smith, H.O. Tomb., J.F. Dougherty, B.A., Fleischmann, R.D. and Venter, J.C. (1995) Science 269, 538–540.

[13] Johnston, M., et al. (1994) Science 265, 2077–2082.

[14] Link, A.J. and Olson, M.V., (1991) Genetics 127, 681–698.

[15] Riles L., Dutchik, J.E., Baktha, A., McCauley, B.K., Thayer, E.C., Leckie, M.P., Braden, V.V., Depke, J.E. and Olson M.Y. (1993) Genetics 134, 81–150.

[16] Williams, N. (1996) Science 272, 481.

[17] Bult, C.J. et al. (1996) Science 273, 1058–1073.

[18] Morell, V. (1996) Science 273, 1043–1045.

[19] Fox, G.E., Magrum, L.J., Balch, W.E., Wolfe, R.S. and Woese, C.R (1977) Proc. Natl. Acad. Sci. USA 74, 4537–4541.

[20] Fox, G.E., Stackebrandt, E., Hespell, R.B., Gibson, J., Maniloff, J., Dyer, T.A., Wolfe, R.S., Balch, W.E., Tanner, R.S., Magrum, L.J., Zablen, L.B., Blakemore, R., Gupta. R., Bonen, L., Leis, B.J., Stahl, D.A., Luehrsen, K.R., Chen, K.N. and Woese, C.R. (1980) Science 209, 457–463.

[21] Woese C., Sogin, M., Stahl, D., Lewis, B.J. and Bonen, L. (1976) J. Mol. Evol. 7, 197–213.

[22] The Flybase Consortium (1994) Nucleic Acids Res. 22, 3456–3458.

[23] Bellen, H.J. and Smith, R.F. (1995) Trends. Genet. 11, 456–457.

[24] Mounkes, L.C., Jones, R.S., Liang, B.-C., Gelbart, W. and Fuller, M.T.A. (1992) Cell 71, 925–937.

[25] Boguski, M.S., Lowe, T.M.J. and Tolstoshev, C.M. (1993) Nat. Genet. 4, 332–333.

[26] Banfi, S., Borsani, G., Rossi, E., Bernard, L., Guffanti, A., Rubboli, F., Marchitello, A., Giglio, S., Coluccia, E., Zollo, M., Zuffardi, O. and Ballabio, A. (1996) Nat. Genet. 13, 167–174.

[27] Hartley, D. (1996) Nat. Genet. 13, 133–134.

[28] Lennon, G.G., Auffrey, C., Polymeropoulos, M. and Soares, M.B. (1996) Genomics 33, 151–152.

[29] Cox D.R., Burmeister, M., Price, E.R., Kim, S. and Myers, R.M. (1990) Science 250, 245–250.

[30] Spence S.E., Gilbert, D.J., Harris, B.S., Bavisson, M.T., Copeland, N.G. and Jenkins, N.A. 1992) Genomics 12, 403–404.

[31] Sulston, J. (1988) in: The Nematode *Caenorhabditis elegans* (Wood, W.B. and the community of *C. elegans* researchers, eds.) pp. 123–155, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

[32] Ellis R.E., Jacobson, D.M. and Horvitz, H.R. (1991) Genetics 129, 79–94.

[33] Hengartner, M.O. and Horvitz, H.R. (1994) Curr. Opin. Genet. Dev. 4, 581–586.

[34] Hengartner, M.O. and Horvitz, H.R. (1994) Cell 76, 665–676.

[35] Hale, A.J., Smith, C.A., Sutherland, L.C., Stoneman, E.A., Longthorne, V.L., Culhane, A.C. and Williams, G.T. (1996) Eur. J. Biochem. 236 1–26.

[36] Brenner, S., Elgar, G., Sandford, R., Macrae, A., Benkatesh, B. and Aparicio, S. (1993) Nature 366, 265–268.