11 Kruglyak, S. and Tang, H. (2000) Regulation of adjacent yeast genes. *Trends Genet.* 16, 109–111
12 Huynen, M.A. and Snel, B. (2000) Gene and context: integrative approaches to genome analysis. In *Analysis of Amino Acid Sequences* (*Adv. Prot. Chem.* Vol. 54) (Bork, P., ed.), pp. 345–379, Academic Press
13 Bork, P. *et al.* (2000) Comparative genome analysis: exploiting the context of genes to infer evolution and predict function. In *Comparative Genomics* (*Computational Biology*) (Sankoff, D. and Nadeau, J.H., eds), pp. 281–294, Kluwer
14 Groth, C. *et al.* (2000) Diversity in organization and the origin of gene orders in the mitochondrial DNA molecules of the genus *Saccharomyces*. *Mol. Biol. Evol.* 17, 1833–1841
15 Blumenthal, T. (1998) Gene clusters and polycistronic transcription in eukaryotes. *BioEssays* 20, 480–487
16 Lathe, W. *et al.* (2000) Gene context conservation of a higher order than operons. *Trends Biochem. Sci.* 25, 474–479

**M.A. Huynen***
**B. Snel**†
**P. Bork**‡
EMBL, Biocomputing,
Meyerhofstrasse 1,
69117 Heidelberg, Germany.
*e-mail: huynen@embl-heidelberg.de
†e-mail: snel@embl-heidelberg.de
‡e-mail: Bork@embl-heidelberg.de

# *Paramecium* genome survey: a pilot project

Philippe Dessen, Marek Zagulski, Robert Gromadka, Helmut Plattner, Roland Kissmehl, Eric Meyer, Mireille Bétermier, Joachim E. Schultz, Jürgen U. Linder, Ronald E. Pearlman, Ching Kung, Jim Forney, Birgit H. Satir, Judith L. Van Houten, Anne-Marie Keller, Marine Froissard, Linda Sperling and Jean Cohen

**A consortium of laboratories undertook a pilot sequencing project to gain insight into the genome of *Paramecium*. Plasmid-end sequencing of DNA fragments from the somatic nucleus together with similarity searches identified 722 potential protein-coding genes. High gene density and uniform small intron size make random sequencing of somatic chromosomes a cost-effective strategy for gene discovery in this organism.**

The ciliated protozoan *Paramecium* was one of the first microorganisms discovered by the early microscopists in the 18th century and has been extensively studied since then. These studies made important discoveries such as microbial sexuality and the occurrence of mating types[1], surface antigens[2], cytoplasmic inheritance[3] and an epigenetic phenomenon not mediated by DNA, called structural heredity[4]. More recently, *Paramecium* has become a powerful model unicell in various fields including membrane excitability[5] and signal transduction[6,7], regulated secretion[8], cellular morphogenesis[9,10], surface antigen variation[11], developmental genome rearrangements[12,13], and homology-dependent epigenetic regulation of both gene expression[14] and developmental genome rearrangements[15]. The recent availability of DNA-mediated transformation[16] allowed complementation cloning of genes identified by mutation[17–19] and gene inactivation by homology-dependent gene silencing through a mechanism related to RNA interference[14,20].

*Paramecium* and the other ciliates are located at a key position in the terminal crown of the eukaryotic phylogenetic tree, together with fungi, plants and metazoa. Moreover, ciliates display a unique feature in the unicellular world: the differentiation of germ and somatic lines in the form of nuclei, not cells. The somatic nucleus (macronucleus) and the germinal nucleus (micronucleus) both derive from the zygotic nucleus, itself derived from parental micronuclei through meiosis and fertilization. During macronuclear development, programmed DNA rearrangements affect the entire genome through amplification to a high ploidy level, chromosome fragmentation and telomere addition, and internal sequence elimination. Many sexual and developmental processes present in metazoa therefore also exist in ciliates, which could serve as pertinent models for their study.

For the moment, no full-scale ciliate genome project has been funded. The community working with *Tetrahymena* has mobilized great ingenuity in genome mapping and development of other tools, including sequencing of expressed sequence tags (ESTs) (J. Fillingham *et al.*, unpublished), with the objective of the complete sequencing of the genome of *Tetrahymena thermophila*[21], a ciliate whose evolutionary distance from *Paramecium tetraurelia* is estimated at greater than 100 Myr.

## The pilot sequencing study

All these considerations stimulated the *Paramecium* community to undertake a genome project. Before being able to establish the full 100–200-megabase genome sequence, *Paramecium* scientists present at the FASEB Ciliate Molecular

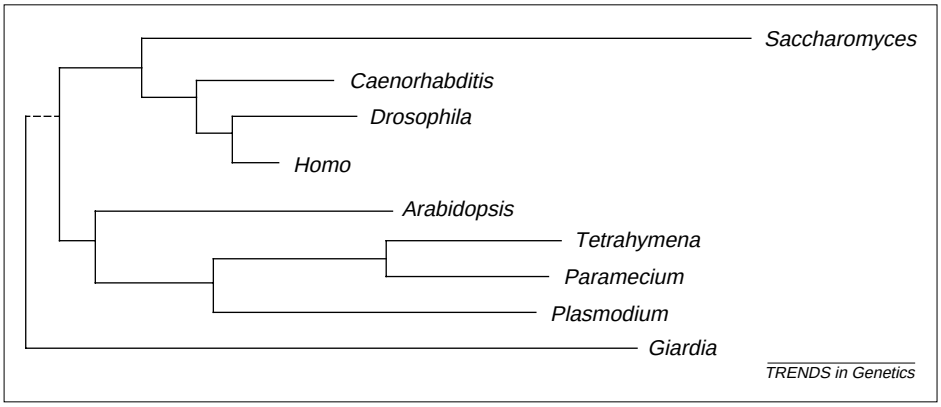


**Fig. 1.** Eukaryotic phylogeny simplified from Ref. 26.

 0168–9525/01/$ – see front matter  PII: S0168-9525(01)02307-1

**Box 1. Identification of potential RNA and protein-coding genes**

**RNA genes**
One tRNA gene
Three rRNA genes

**Protein-coding genes**
722 protein-coding genes (107 303 codons in the partial ORF sequences)
464 introns in 348 of the partial ORF sequences:
256 with one intron
71 with two introns
18 with three introns
3 with four introns
498 of the partial ORF sequences have INTERPRO domains (as identified from the Interpro integrated protein family and domain database; http://www.ebi.ac.uk/interpro/):
472 with one domain
26 with two domains

**Functional classification according to the Munich Information Center for Protein Sequence (MIPS) functional classification** (based on yeast funcat, see http://www.mips.biochem.mpg.de/proj/yeast/catalogues/funcat/index.html)
01 Metabolism: 76 ORFs
02 Energy: 34 ORFs
03 Cell growth: 82 ORFs
04 Transcription: 67 ORFs
05 Protein synthesis: 36 ORFs
06 Protein fate: 84 ORFs
07 Transport facilitation: 57 ORFs
08 Intracellular transport: 64 ORFs
09 Cellular biogenesis: 9 ORFs
10 Signalling: 43 ORFs
11 Cell rescue: 36 ORFs
13 Ionic homeostasis: 33 ORFs
30 Cell organization: 269 ORFs
99 Unknown: 267 ORFs

Biology Meeting (Saxtons River, Vermont, USA; 7–12 August 1999) decided to fund a pilot project of random genomic sequencing on their own grants. The goal of this project was twofold: first, to have an overall idea of the genome organization; and second, to identify as many genes as possible. We decided to take, as random template, the ends of inserts of an indexed genomic library of 6–12-kb macronuclear DNA fragments, initially constructed for complementation cloning[22]. The rationale for this choice was first, that macronuclear genes are active and devoid of intervening sequences characteristic of the micronucleus; and second, that the gene density seems to be very high in the macronucleus (probability of 0.5–0.8 for a base pair to be in a coding sequence). Random sequencing of this library was expected to be almost equivalent to cDNA sequencing, although some noncoding sequences are also present, without the disadvantage of redundant sequencing of highly expressed genes.

Both ends of almost 1800 plasmids were sequenced and 3139 sequences (average length ~500 nucleotides) were obtained after automatic vector screening and concatenation of doublets and contigs. After annotation and removal of very short sequences (<100 nucleotides), 2990 sequence entries containing 1 535 349 nucleotides were submitted to the EMBL/GenBank/DDBJ International Nucleotide database (accession numbers AL446043–AL449029 and AL512551–AL512553).

Analysis of the sequences (L. Sperling *et al.*, unpublished), summarized in Box 1, was performed automatically under a UNIX environment. We used the Phred/Phrap software[23,24] to call bases from the chromatograms, screen vectors and make contigs. Customized Perl scripts were written to control data flow, automatically annotate the sequences and present the data in useful formats such as HTML. The pilot project sequence set was compared with the sequences in the public databases using appropriate BLAST programs[25] to identify homologs of known genes. Visual examination was used to validate the conclusions (quality of the BLAST matches, identification of introns, functional classification). We identified 726 potential genes, among which four specify noncoding RNA (one tRNA and three rRNA) and 722 are protein-coding genes. Only nine of these genes are identical to *Paramecium* genes that were already present in the databases, the remainder are novel genes.

The availability of 722 fragments of potential coding sequences already increases by an order of magnitude the number of *Paramecium* genes in the public databases. The remarkable fact that *Paramecium* uses only small 18–35 base introns (mean 25.4 nucleotides, standard deviation, 2.6 nucleotides) is confirmed and comparison of the open reading frame (ORF) fragments, first identified by the BLAST analysis, to annotated full genomes (*Escherichia coli*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana* and *Homo sapiens*) allowed us to propose a tentative classification into functional categories (Box 1).

We also analysed the sequence similarities of the 448 *Paramecium* ORF fragments that have homologues in each of the completely sequenced eukaryotic genomes. The preliminary result of this analysis is that these conserved *Paramecium* proteins are more similar to human than to worm and fly proteins, and are least similar to yeast proteins. This might seem surprising in view of the known phylogenetic relationships of these organisms (Fig. 1). Because *Paramecium* is a sister group to the clade composed of fungi and metazoans, one would expect equal distances to yeast and animal proteins. However, two factors could explain this result. First, both *C. elegans* and *D. melanogaster* are present on long branches of the

phylogenetic tree[26]. Second, yeast is not only on a long branch, but also lacks (or has lost) cellular structures common to *Paramecium* and metazoans, such as centrioles.

In conclusion, 1.5 megabases of information in 2990 random sequences, representing 1–2% of the *Paramecium* genome, allowed us to identify as many as 722 protein coding genes by homology search. In previous genome projects, a large proportion of genes were found to be species- or phylum-specific 'maverick' genes[27]. It can be anticipated that the sequences in this project contain another 300–500 genes not yet identified owing to their lack of similarity to any sequences presently in databases. The success of the pilot project stimulates us to promote *Paramecium* sequencing further with the aim of getting the full genome soon. All data are available at http://paramecium.cgm.cnrs-gif.fr

**Acknowledgements**

We thank the colleagues in all of our laboratories for their interest and support for this project. We are particularly grateful to André Adoutte, Bernard Dujon and Wlodek Zagorski for useful discussions and encouragement. All of the computing was carried out using the facilities of the INFOBIOGEN Bioinformatics Resource Centre, 91034 Evry Cedex, France.

**References**

1 Sonneborn, T.M. (1937) Sex, sex inheritance and sex determination in *Paramecium aurelia*. *Proc. Natl. Acad. Sci. U. S. A.* 23, 378–385
2 Sonneborn, T.M. (1943) Acquired immunity to a specific antibody and its inheritance in *Paramecium aurelia*. *Proc. Indiana Acad. Sci.* 52, 190–191
3 Sonneborn, T.M. (1943) Gene and cytoplasm: I. The determination and inheritance of the killer characters in *Paramecium aurelia*. II. The bearing of the determination and inheritance of characters in *Paramecium aurelia* on the problems of cytoplasmic inheritance, *Pneumococcus* transformations, mutations and development. *Proc. Natl. Acad. Sci. U. S. A.* 29, 329–343
4 Beisson, J. and Sonneborn, T.M. (1965) Cytoplasmic inheritance of the organization of the cell cortex in *Paramecium aurelia*. *Proc. Natl. Acad. Sci. U. S. A.* 53, 275–282
5 Saimi, Y. and Kung, C. (1987) Behavioral genetics of *Paramecium*. *Annu. Rev. Genet.* 21, 47–65
6 Linder, J.U. *et al.* (1999) Guanylyl cyclases with the topology of mammalian adenylyl cyclases and an N-terminal P-type ATPase-like domain in *Paramecium*, *Tetrahymena* and *Plasmodium*. *EMBO J.* 18, 4222–4232
7 Plattner, H. and Klauke, N. (2001) Calcium in ciliated protozoa: sources, regulation, and calcium-regulated cell functions. *Int. Rev. Cytol.* 201, 115–208
8 Vayssié, L. *et al.* (2000) Molecular genetics of regulated secretion in *Paramecium*. *Biochimie* 82, 269–288
9 Jerka-Dziadosz, M. and Beisson, J. (1990) Genetic approaches to ciliate pattern formation: from self-assembly to morphogenesis. *Trends Genet.* 6, 41–45
10 Beisson, J. and Jerka-Dziadosz, M. (1999) Polarities of the centriolar structure: morphogenetic consequences. *Biol. Cell* 91, 367–378
11 Leeck, C.L. and Forney, J.D. (1996) The 5′ coding region of *Paramecium* surface antigen genes controls mutually exclusive transcription. *Proc. Natl. Acad. Sci. U. S. A.* 93, 2838–2843
12 Caron, F. (1992) A high degree of macronuclear chromosome polymorphism is generated by variable DNA rearrangements in *Paramecium primaurelia* during macronuclear differentiation. *J. Mol. Biol.* 225, 661–678
13 Bétermier, M. *et al.* (2000) Timing of developmentally programmed excision and circularization of *Paramecium* internal eliminated sequences. *Mol. Cell Biol.* 20, 1553–1561
14 Ruiz, F. *et al.* (1998) Homology-dependent gene silencing in *Paramecium*. *Mol. Biol. Cell* 9, 931–943
15 Meyer, E. and Duharcourt, S. (1996) Epigenetic programming of developmental genome rearrangements in ciliates. *Cell* 87, 9–12
16 Meyer, E. and Cohen, J. (1999) *Paramecium* molecular genetics: functional complementation and homology-dependent silencing. *Protist* 150, 11–16
17 Haynes, W.J. *et al.* (1996) Toward cloning genes by complementation in *Paramecium*. *Neurogenetics* 11, 81–98
18 Haynes, W.J. *et al.* (1998) The cloning by complementation of the pawn-A gene in *Paramecium*. *Genetics* 149, 947–957
19 Skouri, F. and Cohen, J. (1997) Genetic approach to regulated exocytosis using functional complementation in *Paramecium*: identification of the ND7 gene required for membrane fusion. *Mol. Biol. Cell* 8, 1063–1071
20 Bastin, P. *et al.* (2001) Genetic interference in protozoa. *Res. Microbiol.* 152, 123–129
21 Orias, E. (2000) Toward sequencing the *Tetrahymena* genome: exploiting the gift of nuclear dimorphism. *J. Eukaryot. Microbiol.* 47, 328–333
22 Keller, A.-M. and Cohen, J. (2000) An indexed genomic library for *Paramecium* complementation cloning. *J. Eukaryot. Microbiol.* 47, 1–6
23 Ewing, E. *et al.* (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessement. *Genome Res.* 8, 175–185
24 Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* 8, 186–194
25 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402
26 Baldauf, S. *et al.* (2000) A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* 290, 972–977
27 Malpertuy, A. *et al.* (2000) Genomic exploration of the *Hemiascomycetous* yeasts: 19. Ascomycetes-specific genes. *FEBS Lett.* 487, 113–121

**P. Dessen**
Service de Bioinformatique,
UMS825 CNRS/SC13 INSERM,
7 rue Guy Môquet BP8, 94801 Villejuif Cedex, France.
**M. Zagulski**
**R. Gromadka**
Institute of Biochemistry and Biophysics,
Polish Academy of Sciences,
DNA Sequencing Laboratory, Pawinskiego 5a,
02-106 Warsaw, Poland.

**H. Plattner**
**R. Kissmehl**
University of Konstanz, Dept of Biology,
78457 Konstanz, Germany.
**E. Meyer**
**M. Bétermier**
Laboratoire de Génétique Moléculaire,
Ecole Normale Supérieure, 46 rue d'Ulm,
75005 Paris, France.

**J.E. Schultz**
**J.U. Linder**
University of Tübingen, School of Pharmacy,
Section Pharmaceutical Biochemistry,
Morgenstelle 8, D-72076 Tübingen, Germany.

**R.E. Pearlman**
Dept of Biology, Core Molecular Biology Facility, York University, 4700 Keele Street,
Toronto, Ontario, Canada M3J 1P3.

**C. Kung**
University of Wisconsin-Madison,
Laboratory of Molecular Biology,
1525 Linden Drive, Madison, WI 53706, USA.

**J. Forney**
Biochemistry Building, Purdue University,
West Lafayette, IN 47907, USA.

**B.H. Satir**
Dept of Anatomy and Structural Biology,
Albert Einstein College of Medicine,
Jack and Pearl Resnik Campus,
1300 Morris Park Ave, Bronx, NY 10461, USA.

**J.L. Van Houten**
University of Vermont, Dept of Biology,
Marsh Life Science Bldg, Burlington,
VT 05405, USA.

**A-M. Keller**
**M. Froissard**
**L. Sperling**
**J. Cohen***
Centre de Génétique Moléculaire,
Centre National de la Recherche Scientifique,
Avenue de la Terrasse,
91198 Gif-sur-Yvette Cedex, France.
*e-mail: cohen@cgm.cnrs-gif.fr