Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

31 Chemin Joseph Aiguier,
13402 Marseille Cedex 20,
France

References

1 Gambello, M.J. and Iglewski, B.H. (1991) *J. Bacteriol.* 173, 3000–3009
2 Passador, L. *et al.* (1993) *Science* 260, 1127–1130
3 Ochsner, U.A. *et al.* (1994) *J. Bacteriol.* 176, 2044–2054
4 Latifi, A. *et al.* (1995) *Mol. Microbiol.* 17, 333–344
5 Jones, S. *et al.* (1993) *EMBO J.* 12, 2477–2482
6 Winson, M.K. *et al.* (1995) *Proc. Natl. Acad. Sci. U. S. A.* 92, 9427–9431
7 Pearson, J.P. *et al.* (1994) *Proc. Natl. Acad. Sci. U. S. A.* 91, 197–201
8 Swift, S. *et al.* (1996) *Trends Biochem. Sci.* 21, 214–219
9 Latifi, A. *et al.* (1996) *Mol. Microbiol.* 21, 1137–1146

Year of the genome

Caroline Ash

Small Genomes: Sequencing, Functional Characterization and Comparative Genomics (TIGR Science Education Foundation conference) held at Hilton Head, SC, USA, 25–28 January 1997.

C. Ash is Editor of Trends in Microbiology.
e-mail: c.ash@elsevier.co.uk

Just 18 months ago there was considerable scepticism about the value being placed on whole genome sequencing efforts. Craig Venter [The Institute for Genomic Research (TIGR), Rockville, MD, USA] described the exponential progress being made, with the imminent publication of the *Escherichia coli* sequence (Fred Blattner, Wisconsin-Madison, USA) and the near completion of the sequences of *Treponema pallidum*, *Borrelia burgdorferi*, (Claire Fraser, TIGR), *Bacillus subtilis* (Antoine Danchin, Institut Pasteur, Paris, France), *Mycobacterium tuberculosis* (Robert Fleischmann, TIGR; Bart Barrell, Sanger Centre, Hinxton, UK) and *Archaeoglobus fulgidus* (Hans-Peter Klenk, TIGR), to name but a few [see Table 1 for an approximation of current genome projects as gleaned from this meeting and various URLs (see Table 2)].

What follows is a necessarily superficial account of the extraordinary volume and diversity of information presented at this meeting. No doubt, during the coming year these data will appear in the public domain at an exhausting rate.

What use are genome sequences?

One example presented at this ground-breaking meeting did elegantly demonstrate the power that knowledge of a whole genome can bring to understanding a biological problem. For several years, Richard Moxon and colleagues (University of Oxford, UK) have been exploring the genetics of phase variation in *Haemophilus influenzae*, the first organism to have its genome completely sequenced[1]. All aspects of the pathogenicity of this microorganism are governed by the properties of its lipopolysaccharide (LPS). The basis for antigenic variation in *H. influenzae* lies in frameshifting among tetrameric repeats that regulate expression of glycosyltransferase and, hence, the biosynthesis of the terminal O antigen region of the LPS molecule. By having access to the whole genome sequence, Moxon's group have exponentially increased their rate of progress and, by probing for known LPS biosynthetic sequences in other organisms, have identified 25 genes scattered around the chromosome. They can now make series of mutants and test their virulence in their infant rat model. Thus, they can map the minimal LPS structural requirements for survival in specific tissues.

How to sequence small genomes

The landmark announcement of the meeting was of the submission of the complete genome sequence of *E. coli* on 16 January 1997 by Fred Blattner and colleagues and the Japan *E. coli* Genome Sequence Group (Hirotada Mori, Nara Institute of Science, Japan)[2]. This project has taken 5 years to sequence ~5 million bases by mapping of restriction fragments and overlapping λ clones.

Shotgun sequencing undoubtedly allows for much faster progress but is reliant on massive computer support. Another essential prerequisite is the preparation of reliably random libraries, something that Ham Smith's team at Johns Hopkins University School of Medicine (Baltimore, MD, USA) excel at. Although ordered cosmid libraries will still be used, for example in the *M. tuberculosis* project, shotgunning is becoming widely adopted and will be used in the *Plasmodium falciparum* genome project (Steve Hoffman, Naval Medical Research Institute, Rockville, MD, USA). Consequently, obtaining 90% of the sequence of a small genome is becoming routine; the other half of the task is to close the gaps between the contigs. This is the point that the *T. pallidum* project (Claire Fraser) has reached.

By now, the *T. pallidum* sequence has probably been closed, and Fraser is pursuing annotation. The next major challenge is to extract meaningful information from the sequence. Already, interesting pointers have come from comparing *T. pallidum* with *Mycoplasma genitalium*. These obligate pathogens

 0966 842X/97/$17.00 PII: S0966-842X(97)01031-7

Table 1. Small prokaryotic and unicellular eukaryotic genome sequence projects as of January 1997[a]

Organism[b]	Participating organizations	Author/contact
Completed and published		
Haemophilus influenzae	The Institute for Genomic Research (TIGR)	Fleischmann[1]
Methanococcus jannaschii	TIGR	Bult[4]
Mycoplasma genitalium	TIGR	Fraser[5]
Saccharomyces cerevisiae	Pasteur consortium	Dujon[3]
Mycoplasma pneumoniae	Heidelberg	Himmelreich[6]
Cyanobacterium synechocystis	Kazusa Research Institute	Kaneko[7]
Completed and unpublished		
Archaeoglobus fulgidus	TIGR	Klenk
Aquifex aeolicus	Recombinant BioCatalysis Inc. (RBI)	Swanson
Borrelia burgdorferi	TIGR/Utah/MedImmune	Fraser/Casjens/Hansen
Escherichia coli K-12	Wisconsin-Madison/Nara, Japan	Blattner/Mori
Helicobacter pylori	TIGR/Genome Therapeutics Corp. (GTC)	Tomb/Smith
Rhizobium NGR234 symbiotic replicon	Jena	Freiberg
Streptococcus pneumoniae	TIGR/MedImmune/GTC	Dougherty
Treponema pallidum	TIGR/U. Texas	Fraser/Weinstock
In progress		
Actinobacillus actinomycetemcomitans	Oklahoma	Roe
Arabidopsis thaliana	Stanford	Federspiel
Aspergillus nidulans	Oklahoma	Kupfer
Bacillus subtilis	Pasteur consortium	Danchin
Caulobacter crescentus	TIGR	?
Chlamydia trachomatis	Stanford/Berkeley	Davis/Stephens
Clostridium acetobutylicum	GTC	Smith
Crenarcheum symbiosum	RBI	Swanson
Deinococcus radiodurans	TIGR	White
Ehrlichia spp.	TIGR	?
Enterococcus faecalis	TIGR/GTC	Dougherty/Smith
Enterococcus faecium	GTC	Smith
Halobacterium salinarium	Max Planck Institute for Biochemistry	?
Legionella pneumophila	TIGR	?
Leishmania major	U. Cambridge	Blackwell
Methanobacterium thermoautotrophicum	GTC/Ohio State	Smith
Mycobacterium avium	TIGR	?
Mycobacterium leprae	Pasteur consortium/GTC/Sanger	Cole/Smith/Barrell
Mycobacterium tuberculosis (various strains)	TIGR/Pasteur/Sanger/GTC	Fleischmann/Cole/Barrell/Smith
Mycoplasma capricolum	George Mason University	?
Neisseria gonorrhoeae	Oklahoma	Roe
Neisseria meningitidis	TIGR/Oxford	Moxon
Neurospora crassa	Oklahoma/New Mexico/Georgia	Kupfer
Plasmodium falciparum chromosomes 1–4,9,10,14	Sanger/NMRI/TIGR	Barrell/Hoffmann
Pseudomonas aeruginosa	Hanover/GTC/U. Washington	Spangenberg/Smith
Pyrococcus furiosus	Utah/Center for Marine Technology	Weiss
Pyrococcus shinkai	MITI/Tokyo	?
Pyrobaculum aerophilum	UCLA/CalTech/RBI	FitzGibbon/Swanson
Rickettsia prowazekii	Uppsala	Andersson
Rhodobacter capsulatus	U. Chicago	?
Sphingomonas aromaticivorans	Dept of Energy	Frazier
Staphylococcus aureus	GTC	Smith
Streptococcus epidermidis	GTC	Smith
Streptococcus pyogenes	Oklahoma	Roe
Sulfolobus solfataricus	National Research Council of Canada consortium	Sensen
Thermoplasma acidophilum	Max Planck Institute for Biochemistry	?
Thermotoga maritima	TIGR	?
Trypanosoma cruzi	Buenos Aires/TIGR	Frasch
Ureaplasma urealyticum	Utah/Alabama	Glass
Vibrio cholerae	TIGR	Clayton
Unresolved status/awaiting funds		
Campylobacter jejuni	UK?	
Brucella abortus	?	
Haemophilus ducreyi	?	
Mycobacterium avium	TIGR	
Pasteurella multocida	?	
Porphyromonas gingivalis	TIGR	
Pyrococcus horikoshii	Japan	
Rhizobium	Toulouse	
Rhodobacter sphaeroides	U. Texas/Houston	
Salmonella typhimurium	TIGR	
Schizosaccharomyces pombe	Sanger	
Shewanella putrefaciens	TIGR	
Streptomyces hygroscopicus	?	
Synechoccus spp.	Japan	
Tetrahymena	?	
Trypanosoma brucei	TIGR	
Treponema denticola	TIGR	

[a]See `http://www.tigr.org` or `http://www.mcs.anl.gov/home/gaasterl/genomes.html`. The latter lists metazoan and virus projects too.
[b]In all, 70 projects are listed here.

Table 2. Some useful URLs[a]

URL	Notes
http://www.tigr.org	Published sequence information, who's doing what where, software, future conferences.
http://www.ai.sri.com/ecocyc/ecocyc.html	*Escherichia coli* database of metabolic pathways.
http://www.er.doe.gov/production/oher/oher_top.html	US Dept of Energy genome project site.
http://www.niaid.gov	National Institute of Allergy and Infectious Diseases site, which includes information about grants for genome initiatives.
http://www.nsf.gov	National Science Foundation site. They are supporting a single genome project for the plant *Arabidopsis thaliana* but the Life in Extreme Environments initiative may be of interest.
http://www.genome.ou.edu	University of Oklahoma site for *Neiserria gonorrhoeae, Streptococcus pyogenes, Aspergillus nidulans, Neurospora crassa* and human mitochondrial genome projects.
http://www.sanger.ac.uk/pathogens	Sanger Centre site for genomes of pathogens.
http://www.imb.nrc.ca/imb/sulfolob/sul	National Research Council of Canada site for the international *Sulfolobolus solfataricus* sequencing effort.
http://www.mcs.anl.gov/home/gaasterl/magpie.html	Terry Gaasterland's MAGPIE project at Argonne National Laboratory.
http://www.mcs.anl.gov/home/gaasterl/genomes.html	Gaasterland's compilation of current genome sequencing efforts.
http://www.ai.sri.com/ecocyc/hincyc.html	*Haemophilus influenzae* metabolic map site.
http://www.atcc.org/hilights/microbe_gene.html	American Type Culture Collection Special Collection of clones resulting from TIGR projects.
http://www.ncbi.nlm.nih.gov	National Center for Biotechnology Information.

[a]These addresses have not all been tested.

show similarities in their parsimonious selection of genes for metabolism, but it is not immediately apparent why *M. genitalium* can grow *in vitro* and *T. pallidum* cannot. A high proportion of the *T. pallidum* genome is devoted to the cell envelope and flagellar genes and, interestingly, generally lacks repetitive elements, except for 15 copies of a major outer sheath protein, which is possibly a porin, analogous to the major outer sheath protein of *Treponema denticola*; Fraser aims to resequence this section to verify its functional characteristics. Fraser is also managing the sequencing of another spirochaete, *B. burgdorferi*, a project that should greatly benefit from the *T. pallidum* sequence. An interesting, but complicating, characteristic of *Borrelia* is the presence of several linear and circular plasmids comprising up to 30% of its bulk DNA. However, these are shed rapidly when the organism is cultured, causing the organism to lose its ability to infect animals, suggesting that the plasmids contain key virulence genes.

Assigning function

Since the completion of the yeast genome project, the impetus of the consortium has refocussed on functional analysis (Bernard Dujon, Institut Pasteur, Paris, France; Elizabeth Winzeler, Stanford University School of Medicine, CA, USA)[3]. The priority is to establish a collection of mutants and complementary clones for the ~5000 genes. Of these genes, 2000 are of mysterious function and appear to be unique to *Saccharomyces cerevisiae*. In two years, a complete repertoire of disruptions should be available. Affymetrix (Santa Clara, CA, USA) have developed a rapid and cost-effective technology for the large-scale parallel analysis of yeast deletion mutants. Their technology enables labelling of each deletion strain with a unique 20-mer tag that can be detected by hybridization to a high-density oligonucleotide array made by their own light-directed oligonucleotide synthesis method. This technology can be potentially extended to monitor expression, by a single hybridization reaction, across the entire genome of selectively grown phenotypes.

There are many pitfalls to assigning function to genes solely on the basis of sequence homology. From her work on the functional analysis of *E. coli*, Monica Riley (Marine Biological Laboratory, Woods Hole, MA, USA) demonstrated the complexity of gene–enzyme relationships: one reaction may depend on multiple isoenzymes (e.g. *fum*) or subunits (e.g. *sdhA–D*) or, conversely, several reactions may only require one gene (e.g. *fadB*). One-to-one relationships are relatively rare (e.g. *speD*). Riley has constructed dendrograms of enzyme specificity to trace evolutionary lineages, map sequence to activities and reveal cryptic relationships, all of which will provide valuable clues for the annotation of other genomes. Riley estimates that about 15% of the genes in *E. coli* are enzymes, 14%

have a transport function, 12% are regulators, 6% are associated with RNA and 5% are carriers, structural proteins and other factors. The large number of coding regions with unknown function are known as orphan genes, and methods to pinpoint their roles generated much debate and speculation.

One development of Riley's strategy is 'computational metabolism'. Peter Karp (SRI International, Menlo Park, CA, USA) is mapping the metabolic pathways of *H. influenzae* and *E. coli* (Table 2). Metabolic function in *E. coli* is well worked out (730 enzymes in 125 pathways and 3020 genes) and, together with its sequence information, forms an excellent database to extrapolate function to genes in related species. Karp has mapped 81 pathways in *E. coli*, of which 60 may also exist in *H. influenzae*, although information on cofactors is lacking (372 reactions in 73 pathways and 1746 genes). This work has revealed, for example, that purine biosynthesis in *H. influenzae* appears to be identical to that in *E. coli*, but that *H. influenzae* does not have a complete glycolytic pathway. One important facet of this work is that it can be used to inform sequence analysis with reference to the biology of the organism.

Comments were made about the lack of biological relevance in Eugene Koonin's (National Center for Biotechnology Information, Bethesda, MD, USA) highly theoretical ortholog and paralog computational comparisons of eubacterial and archaeal genomes. Nevertheless, such computational exercises can help recognition of what comprises the complete genomic repertoire of a microorganism and the identification of functions that are truly lacking or are novel. It is perhaps at the level of fine detail that this methodology becomes meaningless unless the lifestyle of the organism is taken into account. From his analysis, Koonin estimates a sequence homology (at the 70% level) across all the known bacterial and archaeal genomes of 35–40% and is well known for his estimate of the theoretical minimal genome (~250 genes).

Janet Seifert (University of Houston, TX, USA) highlighted the important observation that gene order is remarkably poorly preserved across the gamut of bacteria. She estimates that there are only 19 gene clusters with a common organization across the four known eubacterial genomes and only seven, if the comparison is extended to the archaeans. Of these, five clusters include the most ancient genes (tRNA), which apparently became coordinated early on in evolution. She has grouped essential genes into those required for the translation machinery, transcription, ABC and SecE transport, ATP synthase, chaperones, including GroEl and GroES, and polyamine transport. Her results are consistent with the notion that ancestral genomes were RNA and contained several genes.

The mycoplasmas offer some insight into what makes a minimal genome. *M. genitalium* has a genome of 0.58 Mb, which contains ~500 genes. Clyde Hutchison (University of North Carolina, Chapel Hill, USA) is using the *E. coli* pathway maps to define the metabolic needs of *M. genitalium*. It appears to have no amino acid biosynthetic pathways and the search is on for corresponding amino acid or peptide transporters. From gene knockout experiments, Hutchison suggests that *M. genitalium* could remain viable with a minimum of 400 genes under defined culture conditions, which compares with Muschegian and Koonin's estimate for the primordial bacterial genome of 256 genes. Interestingly, the related pathogen *Mycoplasma pneumoniae* has ~200 more genes than *M. genitalium* but it is not known why. Nevertheless, an obligate pathogen cannot provide direct clues for what constitutes a minimal gene complement for self-sustaining life. The mycoplasmas have apparently evolved to become obligate pathogens from Gram-positive bacteria like *Spiroplasma*. Richard Herrmann (University of Heidelberg, Germany) wants a third mycoplasma sequence to resolve what actually is ancestral among the extensive gene rearrangements seen between the two sequenced species.

Rickettsia prowazekii is also an obligate intracellular pathogen with a small (1.1 Mb) and plastic genome that shares some relationship with the mammalian mitochondrial genome. *R. prowazekii* has few amino acid biosynthetic pathways but, unlike *M. genitalium*, which relies on glycolysis alone, *R. prowazekii* only requires a tricarboxylic acid cycle to generate energy (Siv Andersson, Uppsala, Sweden). The main genetic problem for this organism is that it has a semi-clonal population structure. Andersson has been monitoring strain variation in S-adenosylmethionine synthetase, an essential enzyme for polyamine metabolism. It suffers a very high rate of deletion and susbstitution with a tenfold greater fixation rate than the corresponding enzyme of *E. coli*; a feature that Andersson expects to be common for intracellularly replicating parasites. The end result for any particular clone is mutational meltdown. This organism fulfils Maynard-Smith's maxim that life without sex is an evolutionary scandal.

Wild genomes

Although the majority of small genomes that have been, or are being, sequenced are of pathogens, there is certainly an enthusiastic gold rush feeling about exploring genomes obtained from other environments. Regardless of the potential for novel industrial products, Norman Pace (University of California, Berkeley, USA) is fervent in his exploration for uncultivatable organisms that represent the spectacular majority of biodiversity. Pace uses conserved rRNA-based PCR primers to derive 'phylotypes' of organisms. His protégé, Ed deLong (University of California, Santa Barbara, USA) has taken on the challenge by monitoring seasonal distributions of uncultivatable archaeans in oceanic waters. The metabolic characteristics of such psychrophiles are not so dissimilar from those of hyperthermophile archaeans like *A. fulgidus*, which grows at 64–92°C and uses simple organic substrates to generate energy. Genomic comparison should distinguish the requirements for high-temperature adaptation.

The 2.2-Mb *A. fulgidus* sequence is probably complete by now (Hans-Peter Klenk) and annotation is under way. Only a few genes of known function have been firmly identified, but it appears to share up to half the genes of *Methanococcus jannaschii* although in no organizational relationship. One apparent consequence of life at high temperature is to have more and shorter open reading frames.

Another way of sampling the environment for potentially exploitable microorganisms is simply to screen for gene products with no reference to the originating genomes. Eric Mathur (Recombinant BioCatalysis Inc., La Jolla, CA, USA) outlined their strategy for the discovery of industrial enzymes from uncultivatable organisms; they use a combination of phylogenetic trees as 'road maps' and recombinant protein screening from 'environmental libraries' obtained from DNA extracted directly from samples. His company is sampling the extremes, from insect guts to whale carcasses (which contain microorganisms with interesting esterases), as well as the, by now obligatory, survey of hydrothermal vent flora; the only bottleneck to their environment-wide surveying is the limitations of their robotic screening capacity. They are looking for enzymes with specific properties, such as thermal tolerance, long shelf life and resistance to solvents, and so far have recovered 300 industrially relevant enzymes.

Conclusions

There is still a sense of gaining immense power from genomics but having little to fire it at, as knowledge of an entire genome will not automatically answer a lot of questions about an organism's biology. Of course, it will tell us general features (Robert Fleischmann) about small genome organization, pinpoint many of the open reading frames and open the door to the discovery of genes for virulence, drug and vaccine targets, diagnostic tests and industrially important enzymes, among many other properties. What remains a limitation is a lack of agreement about gene nomenclature and the huge potential for confusion and misattribution of function (Craig Venter). But the main thing that knowledge of the genome does reveal is how little we do actually know and the potentially astonishing complexity and plasticity of even the smallest genomes. It also demonstrates the quite extraordinary ingenuity and collaborative powers of the disparate group of technologists who work on genome sequencing.

The next meeting, Small Genomes II, will also be held at Hilton Head Island from 31 January to 4 February 1998. See `genome@tigr.org` for details.

References

1 Fleischmann, R. *et al.* (1995) *Science* 269, 496–512
2 O'Brien, C. (1997) *Nature* 385, 472
3 Dujon, B. (1996) *Trends Genet.* 12, 263–270
4 Bult, C.J. *et al.* (1996) *Science* 273, 1058–1073
5 Fraser, C. *et al.* (1995) *Science* 270, 397–403
6 Himmelreich, R. *et al.* (1996) *Nucleic Acids Res.* 24, 4420–4449
7 Kaneko, T. *et al.* (1996) *DNA Res.* 3, 109–136

Letters

A paradigm for virus-induced demyelinating disease

Paradigm: an outstandingly clear and typical example.
Longmans Dictionary

In their recent review in *Trends in Microbiology*[1], Lane and Buchmeier present murine infection with the mouse hepatitis coronavirus (MHV) as a paradigm for virus-induced demyelinating disease in humans, in particular for multiple sclerosis (MS). Such a paradigm would presumably also be valid for the known human virus-induced demyelinating diseases such as postmeasles encephalomyelitis, progressive multifocal leuko-encephalopathy (papovavirus) or tropical spastic paraparesis (human T cell leukaemia virus type 1).

The etiology of the multifactorial disease MS is still unknown, but MS remains a prime candidate for a virus-induced demyelinating disease[2], and an adequate animal model would certainly be a great benefit to MS research. A series of inbred murine (or rodent) models for MS exist: an antigen-based experimental allergic encephalomyelitis (EAE) and several virus-mediated models, including coronavirus, lactic dehydrogenase/N-tropic C-type virus[3] and Theiler's encephalomyelitis virus[4]. All these models suffer from two major drawbacks: they do not represent the highly variable manifestations of MS (from subclinical course over relapsing and remitting forms to slow or fast chronic progressing disease) and they are all rodent-based. The basic structure of the immune system in rodents and humans is very similar and, therefore, it is tempting to assume that an immunoregulatory mechanism found in one is automatically present in the other. This is indeed the case in some instances, but not always: a costly lesson that was learned in MS therapy[5].

It seems reasonable to caution against presentations like Table 1 in the review by Lane and Buchmeier. An MS exacerbation does not necessarily correspond to an MHV-induced exacerbation in mice, and the implications of a cytokine response differ in the two cases. We still do not have a paradigm for human virus-induced demyelinating diseases/MS, and mice are not men.

Tove Christensen
Dept of Medical Microbiology and Immunology, Bartholin Building, University of Aarhus, DK-8000 Aarhus C, Denmark

References

1 Lane, T.E. and Buchmeier, M.J. (1997) *Trends Microbiol.* 5, 9–14
2 Christensen, T., Møller-Larsen, A. and Haahr, S. (1994) *Trends Microbiol.* 2, 332–336
3 Pease, L.R. and Murphy, W.H. (1980) *Nature* 286, 398–400
4 Dal Canto, M.C. and Rabinowitz, S.G. (1982) *Ann. Neurol.* 11, 109–127
5 Biliau, A. (1995) *Neurology* 45, 550–553