More Haemophilus and Mycoplasma Genes

Jurgen Brosius; Keith Robison; Walter Gilbert; George M. Church; J. Craig Venter

Science, New Series, Vol. 271, No. 5253 (Mar. 1, 1996), 1302-1304.

Stable URL:
http://links.jstor.org/sici?sici=0036-8075%2819960301%293%3A271%3A5253%3C1302%3AMHAMG%3E2.0.CO%3B2-%23

(9) regard this as a "time-sensitive" species, its previous identification at only two sites, which on his analysis are either synchronic or separated by no more than 0.1 to 0.2 Ma, means that it is likely that only a part of this species' range in time has so far been sampled. It cannot thus have any bearing on the dating of Member 2. Its absence from Makapansgat cannot be used as dating evidence, for it may indicate taphonomic or environmental factors, a principle which McKee and others recognized.

Phillip V. Tobias
Ronald J. Clarke
Paleo-anthropology Research Unit, Department of Anatomical Sciences, University of the Witwatersrand Medical School, Parktown, Johannesburg 2193, Republic of South Africa

REFERENCES

1. T. C. Partridge, *Nature* **275**, 283 (1978).
2. D. N. Stiles and T. C. Partridge, *S. Afr. J. Sci.* **75**, 346 (1979).
3. T. C. Partridge and I. B. Watt, *Palaeontol. Afr.* **28**, 35 (1991).
4. R. J. Clarke and P. V. Tobias, *Science* **269**, 521 (1995).
5. P. L. McFadden *et al.*, *Earth Planet. Sci. Lett.* **55**, 373 (1979).
6. T. C. Partridge, personal communication.
7. F. J. Hilgen, *Earth Planet. Sci. Let.* **107**, 349 (1991).
8. T. D. White *et al.*, *Nature* **371**, 306 (1994).
9. J. K. McKee *et al.*, *Am. J. Phys. Anthropol.* **96**, 235 (1995).

5 December 1995; accepted 6 February 1996

More *Haemophilus* and *Mycoplasma* Genes

The sequencing of two entire genomes of free living organisms by Fleischmann *et al.* and by Fraser *et al.* is an outstanding achievement (*1*) and provides a rich basis for further studies. The methods used in these studies to detect genes, however, are biased toward protein coding genes and ribosomal and transfer RNAs. These methods do not reflect the significance of additional RNA species that vitally contribute to the functioning of a cell. In the 0.54- and 1.83-megabase *Mycoplasma genitalium* and *Haemophilus influenzae* genomes, several small stable RNA species remained undetected (*1*): I have located (*2*) the M. *genitalium* and *H. influenzae* genes encoding 4.5S RNA [part of the SRP homolog of Bacteria (*3*)], ribonuclease (RNase) P RNA (*4*), 10*Sa* RNA, and the *H. influenzae* 6S RNA (Table 1). The cellular roles of the latter two RNAs are still unknown [10*Sa* RNA may be involved in the COOH-terminal extension of protein (*5*)]. The proposed half-tRNA–like structure for a portion of 10*Sa* RNA (*6*) is conserved and further supported by compensating base changes in the *M. genitalium* and *H. influenzae* orthologs. In M. *genitalium*, the RNase P RNA gene is tightly linked with the 10*Sa* RNA gene (divergently transcribed), while in *H. influenzae*, the two genes are about 400 kb apart (*7*). The two RNA genes are so close in proximity that their putative −10 promoter regions (TATAAT) overlap by four nucleotides. This may imply a concerted gene regulation.

Very likely, there are more hidden RNA treasures in the genomes of *M. genitalium* and *H. influenzae*. Analysis of the many gaps between open reading frames (ORFs) may reveal undiscovered functional RNA species.

Jürgen Brosius
Institute for Experimental Pathology, (Zentrum für Molekularbiologie der Entzündung), University of Münster, Von-Esmarch-Strasse 56, D-48149 Münster, Germany
E-mail: rna.world@uni-muenster.de

Table 1. RNA species detected in *M. genitalium* and *H. influenzae*. RNA sequences are available on the World Wide Web at URL <http://www-ifi.uni-muenster.de/xpath/mge-hin-rna.html>.

RNA	Bacterial species*	Query sequence†	Position‡
4.5*S*	Mge	Mpn (S76009)	325,924–326,002§
	Hin	Eco (X01074)	1,223,295–1,223,408
6*S*	Mge	Eco, Hin, Pae (M12965, U32767, and Y00334)	None‖
	Hin	Pae (Y00334)	905,485–905,683
10*Sa*	Mge	Mca (D13067)	406,541–406,928
	Hin	Eco (D12501)	1,356,974–1,357,339§
RNase P	Mge	Mhy (X69982)	406,137–406,518§
	Hin	Eco (U18997)	1,745,841–1,746,216§

*Abbreviations: Eco, *Escherichia coli*; Hin, *Haemophilus influenzae*; Mca, *Mycoplasma capricolum*; Mge, *Mycoplasma genitalium*; Mhy, *Mycoplasma hyopneumoniae*; Mpn, *Mycoplasma pneumoniae*; and Pae, *Pseudomonas aeruginosa*. †Searched with the accession number listed. ‡The exact 5′ and 3′ ends of the mature *M. genitalium* and *H. influenzae* RNAs were not experimentally determined and are inferred from related RNA sequences and secondary structures. §RNA gene is complementary with respect to the numbered sequence. ‖No match was found with the search program applied.

REFERENCES AND NOTES

1. R. D. Fleischmann *et al.*, *Science* **269**, 496 (1995); C. M. Fraser *et al.*, *ibid.* **270**, 397 (1995).
2. The BLASTN search program [S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990)] was applied for the *H. influenzae* genome and the GRASTA program (*1*) for *M. genitalium*.
3. M. A. Poritz *et al.*, *Science* **250**, 1111 (1990).
4. C. Guerrier-Takada, K. Gardiner, T. Marsh, N. R. Pace, S. Altman, *Cell* **35**, 849 (1983).
5. G. F. Tu, G. E. Reid, J. G. Zhang, R. L. Moritz, R. J. Simpson, *J. Biol. Chem.* **270**, 9322 (1995).
6. Y. Komine, M. Kitabatake, T. Yokogawa, K. Nishikawa, H. Inokuchi, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 9223 (1994); C. Ushida, H. Himeno, T. Watanabe, A. Muto, *Nucleic Acids Res.* **22**, 3392 (1994).
7. In *M. genitalium*, the 10*Sa* gene is also adjacent to an ORF that has putatively been identified as aminopeptidase P (pepP). In *H. influenzae*, the 10*Sa* gene is in close proximity to $tRNA^{Met}$-C, nusA, and infB. In *Thermus thermophilus*, a portion of the 10*Sa* RNA gene (accession number Z48001, pos. ~220-1) has been sequenced; it is also closely linked to the nusA/infB operon. HI0858 ORF (adjacent to the 6*S* RNA gene) has sequence similarity to the *E. coli* ORF located downstream of 6S RNA.

19 September 1995; accepted 3 November 1995

The publication of the complete genomic sequences for *H. influenzae* by Robert D. Fleischmann *et al.* (*1*) and M. *genitalium* by Claire M. Fraser *et al.* (*2*) presents the first opportunities to compile complete catalogs of proteins for free-living organisms. It is important that such catalogs be as comprehensive as possible, especially in the case of M. *genitalium*, which is believed to define a "minimal gene content" for a bacterium (*2*). Refinement of the descriptions within the *H. influenza* gene catalog in the study by Fleischmann *et al.* (*1*) has already been attempted by Casari *et al.* (*3*). However, critical examination of these catalogs is an essential prerequisite to their further analysis, as they are incomplete.

We have found an additional 17 protein-coding regions beyond the 1743 reported for *H. influenza* and 3 in addition to the 470 reported for M. *genitalium* using our previously described (*3*) database search strategy (Table 1). These new genes range from proteins ubiquitous to life (three ribosomal proteins, a cold-shock domain-type DNA binding protein, and a DNA repair enzyme) to genes similar only to proteins of unknown function.

In several cases, the close proximity of these new genes to those previously reported or to genes of related function suggests operon structures. For example, in M. *genitalium* the DNA repair enzyme fpg (M2) is preceded by DNA polymerase genes (MG261 and MG262), and the atpB (M4) gene is embedded among nine other genes for the adenosine triphosphatase (ATPase) complex (MG405-

Table 1. New genes found by similarity search. All regions of *Haemophilus* (*1*) and *Mycoplasma* (*2*) not reported in the respective papers as protein-coding in HIDB or MGDB were searched with the use of BLASTX against the combined SwissProt+PIR+GenBank translations database with the use of the NCBI Network BLAST server (*4, 5*) and with the use of the BLOSUM62 matrix, with a scoring cutoff of 60. After elimination of matches reported in HIDB and MGDB, the ORFs corresponding to each match were extracted. Results suggesting extensions of ORFs in HIDB and MGDB were then identified (Table 2). The first in-frame ATG codon was used as the start codon, unless similarity was observed upstream, in which case the first NTG codon upstream of the observed similarity was used. All of the ORFs listed have BLASTP matches with Poisson probability estimates $<10^{-8}$. For each ORF is listed the location of the start and ending position on the *Haemophilus* genome, the flanking ORFs from HIDB/MGDB, length in amino acids, and description. HIDB/MGDB numbers marked (*) indicate probable operon relationships, determined on the basis of tandem ORFs either participating in the same biochemical pathway or in close proximity of the start and stop codons of the two ORFs. When available, the name for the most similar *E. coli* protein is provided within parentheses. Further elaboration of these results is accessible over the World Wide Web at URL <http://golgi.harvard.edu/hinmge/neworfs/table1.html>.

ORF	From	To	Flanking ORFs		Size	Description
			New Haemophilus influenzae *genes*			
H1	169988	168807	HI0152*	HI0153	394	C4-dicarboxylate transporter (dcuB)
H2	208082	211417	HI0194	HI0196*	1112	Similar to 2 ORFs from *E. coli* (yjeP & yggB)
H3	248228	247320	HI0220	HI0221	302	Biotin acetyl-CoA carboxylase/biotin repressor (birA)
H4	409674	411362	HI0389	HI0390*	563	Long-chain fatty acid acetyl-CoA ligase (lcfA)
H5	595012	595479	HI0576*	HI0577*	119	Similar to *E. coli* ORF (yrdD)
H6	654418	653774	HI0620*	HI0621	215	Transport protein (yaeE)
H7	655621	656175	HI0621	HI0622	185	Imidazoleglycerol-phosphate dehydratase–like ORF (yaeD)
H8	793969	795510	HI0738*	HI0739	514	Threonine dehydratase (ilvA)
H9	849799	849912	HI0798*	HI0799*	38	Ribosomal protein L36 (rpmJ)
H10	1007576	1006041	HI0946	HI0947	512	L-2,4-diaminobutyrate decarboxylase
H11	1032176	1032433	HI0974	HI0975*	86	Similar to *E. coli* ORF (yhdT)
H12	1295045	1294506	HI1225	HI1227	228	dnaA isolog (dnaA)
H13	1485614	1487047	HI1389*	HI1390	478	Indole-3-glycerol phosphate synthase (trpC)
H14	1522623	1522402	HI1434	HI1435	74	Cold shock-type DNA binding protein (cspD)
H15	1592546	1592785	HI1522	HI1523	40	Phage Mu Com protein
H16	1822923	1821778	HI1739	HI1740	382	L-lactate dehydrogenase (lctD)
H17	1823893	1823135	HI1739	HI1740	253	Glutamate racemase (murI)
			New Mycoplasma genitalium *genes*			
M1	64514	64368	MG055	MG056	49	Ribosomal protein L33 (rpmG)
M2	319205	320059	MG262*	MG263	284	Formamidopyridine-DNA glycosylase (mutM)
M3	460951	460685	MG363	MG364	89	Ribosomal protein S20 (rpsT)

MG397). Ribosomal protein S20 (M3) is clustered with ribosomal proteins L10, L7/L12, and L32 (MG361, MG362, and MG363, respectively), although S20 is in the opposite orientation from the other three. Similarly, in *H. influenza*, ribosomal protein L36 (H9) is embedded in a large cluster of ribosomal proteins (HI0776-HI0803), ilvA (H8) is downstream of other isoleucine-leucine-valine biosynthesis genes, and trpC (H13) is located downstream of the tryptophan biosynthesis genes trpE (HI1387) and trpD (HI1388 and HI1389). Protein ORF H5 is present in an operon (HI05577-H5-HI0576-HI0575), of which the last three genes are present in the same arrangement in *Escherichia coli*; none of these in either species has an identified function.

In addition, we found that six of the ORFs predicted by Fleischmann *et al.* and nine predicted by Fraser *et al.* can be significantly extended at the amino terminus with the use of similarity as a guide to finding another eubacterial start codon (Table 2). In two cases, extension required the introduction of a frameshift. For example, we initially identified a homolog of MG414/MG413 (which are two versions of the same ORF). Because the region of similarity spanned the stop codon separating our ORF from MG415, we suggest that MG415 has suffered a mutation. Both MG415 and MG414 are large (664 and 1036 amino acids, respectively). The similarity between them covers almost the entire length of MG415, but neither shows extensive similarity to other database proteins. Also, they are located adjacent to one another in tandem orientation.

The search for protein-coding regions in these genomes is probably not over. Sequencing of other microbial genomes is likely to suggest more ORFs by similarity (*4, 5*), which together with careful study of operon structures may reveal short reading frames that have thus far escaped detection.

Keith Robison*
Walter Gilbert
Department of Molecular and Cellular Biology, Harvard Biological Laboratories, 16 Divinity Avenue, Cambridge, MA 02138, USA
**E-mail: krobison@nucleus.harvard.edu*
E-mail: gilbert@nucleus.harvard.edu
George M. Church
Department of Genetics, Harvard Medical School, Warren Alpert Building, 200 Longwood Avenue, Boston, MA 02115, USA
E-mail: church@rascal.med.harvard.edu

REFERENCES

1. R. D. Fleischmann *et al.*, *Science* **269**, 496 (1995).
2. C. M. Fraser *et al.*, *ibid.*, **270**, 397 (1995).
3. G. Casari *et al.*, *Nature* **376**, 647 (1995).
4. K. Robison, W. Gilbert, G. M. Church, *Nature Genetics* **7**, 205 (1994).
5. S. F. Altschul, M. S. Boguski, W. Gish, J. C. Wootton, *ibid.*, **6**, 119 (1994).

10 October 1995; revised 1 November 1995; accepted 3 November 1995

Response: The recent publications by our laboratory of the first two completely sequenced genomes from free-living organisms [Fleischmann *et al.* (*1*) and Fraser *et al.* (*2*)] have laid the groundwork for in-depth biological analysis of these organisms. Our initial analysis of these genomes, using computational methods, was deliberately conservative, and our publications reflect the limited conclusions that can be drawn on the basis of such methods. Our assignment of functional roles was meant to suggest a set of functional hypotheses based on sequence similarity analysis. These alignments are provided on our Web site (at URL <www.tigr.org>), which contains our annotation and updates of the *H. influenzae* Rd and *M. genitalium* genomes and has allowed the scientific community to participate in the continued analysis of these two genomes. We have received dozens of valuable comments and suggestions through the Web site. A comment by Casari *et al.* (*3*) has appeared in print with respect to the *H. influenzae* and *M. genitalium* analysis.

The identification by Brosius of four stable RNA species (4.5S, M1, 10Sa, and 6S) in the *H. influenzae* genome and three stable RNA species (4.5S, 10Sa, and M1) in *M. genitalium* is a valuable contribution to the continued analysis and updating of the genome annotation and illustrates the importance of developing comprehensive

Table 2. Open reading frames reported by Fleischmann *et al.* (*1*) or Fraser *et al.* (*2*) that can be extended at the amino terminus on the basis of similarity. The first NTG codon upstream of the observed similarity was chosen. The HIDB and MGDB identifier, start codon sequence, length of extension in amino acids, percent increase in ORF size, and description of the database match are listed. The extensions of the entries marked (*) require the introduction of a single frameshift each. The entry marked $ (HI1477) is described as frameshifted in HIDB, but we find no evidence for a frameshift. Further elaboration of these results is accessible over the World Wide Web at URL <http://golgi.harvard.edu/hinmge/neworfs/table2.html>.

Start nt	ORF	Start codon +aa	+%	Description
		Extended Haemophilus *ORFs*		
620821	HI0599	TTG	25	19	Possible regulatory protein (recX)
799267	HI0740	ATG	65	13	Phosphomannomutase (ureD)
1023925	HI0965	TTG	26	42	Ribosomal protein S20 (rpsT)
1228744	HI1159	TTG	26	12	Thioredoxin (trxA)
1483506	HI1465	ATG*	78	67	Anthranilate synthase component I (trpD)
1560321	HI1477	ATG$	27	44	DNA binding protein (ner)
		Extended Mycoplasma *ORFs*		
20550	MG018	TTG	126	30	Probable DNA repair helicase
39128	MG033	TTG	38	17	Glycerol uptake facilitator (glpF)
49829	MG042	TTG	73	15	Putrescine transporter (potG)
142080	MG115	TTG	78	99	Family of ORFs (ygaD)
245632	MG206	TTG	165	38	DNA repair enzyme (uvrC)
392257	MG313	TTG	78	37	Cytoadherance accessory protein hmw1
504258	MG399		94	25	ATPase component (atpD)
521414	MG415	ATG*	392	144	*Mycoplasma* ORF MG414
527014	MG421	TTG	106	13	DNA repair enzyme (uvrA)

analysis tools that will aid in identifying not only genes encoding proteins but also genes encoding RNA elements and promoter and regulatory regions.

Robison *et al.* comment on two aspects of genome analysis: (i) the identification of additional protein coding regions in the "intergenic" regions of *H. influenzae* and *M. genitalium* (Table 1) and (ii) the upstream extension of open reading frames. Our own prior analysis of the "intergenic" regions in *H. influenzae* resulted in the identification of 15 of these coding regions as well as 5 additional ones. These regions have been updated on our Web site and submitted to GSDB.

Over the past few months we have refined our identification of start codons for genes. In particular, we have examined overlapping start and stop codons from adjacent genes and resolved conflicts where possible. These changes are reflected on our Web site and GSDB. However, many documented cases remain where start and stop codons of adjacent genes are either overlapping or partially shared, thus we have altered the start codon only in those cases where there is well-documented knowledge as to the physiologically relevant start codon.

Casari *et al.* (*3*) have given the results of applying the GeneQuiz analysis system to the predicted coding regions of the *H. influenzae* genome for which only a hypothetical database match or no database match were previously reported. Annotation with this completely automated system resulted in 148 putative protein functions being reported, and the results are said to be as exhaustive and accurate as those reached by sequence analysis experts. The goals associated with our annotation of the *H. influenzae* genome sequence and those of the GeneQuiz approach are in some ways quite different, and it is important to clarify the differences in the annotation provided by Casari *et al.* and that presented by us in Fleischmann *et al.* Our intent was to define a predicted coding region according to its best database match, taking a conservative approach to associating a function on the basis of shared sequence similarity. Casari *et al.* (*3*) have attempted to identify potential protein family relationships on the basis of sequence similarity and conserved signatures. The danger of assigning protein functions on the basis of conserved domain structures in a fully automated fashion is illustrated by HI0017. We agree with Casari *et al.* (*3*) that the best database match is to SP:P33633, a hypothetical protein that may be activated under anaerobic conditions through the formation of an organic free radical at a conserved glycine residue. Enzymes in this family are inactivated under aerobic conditions by cleavage of the peptide at the glycine residue harboring the organic radical. Casari *et al.* (*3*) have assigned a functional role of formate acetyltransferase 1 to HI0017 on the basis of an alignment to SP:P09373. However, there is no shared alignment in the region associated with the catalytic activity of formate acetyltransferase 1 (the conversion of acetyl coenzyme A + formate to pyruvate + coenzyme A). The 92.9% shared similarity of predicted coding region HI0180 with the *E. coli* formate acetyltransferase 1 across its entire length suggests that HI0180 is the most likely homolog of formate acetyltransferase 1 in *H. influenzae* (*1*).

As more complete microbial genome sequence data become available in the coming years and the sequencing of the human genome begins in earnest, the flow of genomic data will dramatically increase. The development of systems that can analyze those data in an automated fashion should be a goal of laboratories participating in genome sequence and analysis endeavors. Automated gene analysis systems need to incorporate a variety of information, such as confidence values, multiple sequence alignments, and sequence context in order to develop functional hypotheses from genome sequence information. However, only laboratory experiments will ultimately resolve the biological questions raised by the availability of these data.

J. Craig Venter
Institute for Genomic Research (TIGR), 9712 Medical Center Drive, Rockville, MD 20850, USA

REFERENCES

1. R. D. Fleischmann *et al.*, *Science* **269**, 496 (1995).
2. C. M. Fraser *et al.*, *ibid.* **270**, 397 (1995).
3. G. Casari *et al.*, *Nature* **376**, 647 (1995).

10 October 1995; accepted 3 November 1995