correspondence between case reports and fatality data. These data also establish that mortality rates are not affected by epidemic phase[24]. Further confirmation of these results is provided by an analysis of the Aberdeen data (N.B.M-B., P.R. and B.T.G., manuscript in preparation). Concerning infection-induced mortality rates, classic work by Butler[24], Bartlett[25], Creighton[5] and others indicates significant mortality due to measles and whooping cough during these periods. Estimates of case fatality rates for measles vary widely, from 1–2% in the post-war era up to 46% pre-war[14,26,27], whereas estimates for whooping cough are in the 3–15% range[24].

Data analysis

These time series contain a substantial annual component and are further complicated by increasing population sizes over the two periods examined. Hence, analyses of the relationship between measles and whooping cough outbreaks were carried out on de-trended data. We used three separate methods. First, Pearson correlation coefficients were estimated for data aggregated over each epidemic year (October to October). Second, we carried out a linear regression of annual counts of measles against whooping cough and used the slope as a measure of synchrony. The results of this technique were qualitatively identical to those of the Pearson correlation, so we present only those. Finally, we also used Wavelet spectra to explore phase differences between filtered time series[28,29]. Further information can be found in the Supplementary Information.

Bifurcation analysis

The bifurcation diagram was prepared by solving the system of ordinary differential equations (ODEs) for each parameter combination and characterizing the dynamics after a 190-year transient period was ignored. We used wrap-around initial conditions (whereby the final conditions for a run were used as the initial conditions for the next run). In the regions where the birth rates are relatively small, more than one attractor can coexist, although annual dynamics have the largest basin[8]. Note that in model dynamics, for some very restricted choice of initial conditions, it is possible to observe biennial dynamics that are in phase. However, work in progress has shown that infection dynamics become out of phase when stochasticity is introduced into the model, consistent with the findings of Kamo and Sasaki[15].

Received 9 January; accepted 25 February 2003; doi:10.1038/nature01542.

1. Gog, J. R. & Swinton, J. A. A status-based approach to multiple strain dynamics. *J. Math. Biol.* **44,** 169–184 (2002).
2. Gupta, S., Ferguson, N. M. & Anderson, R. M. Chaos, persistence and evolution of strain structure in antigenically diverse infectious agents. *Science* **280,** 912–915 (1998).
3. Gomes, M. G. M., Medley, G. F. & Nokes, D. J. On the determinants of population structure in antigenically diverse pathogens. *Proc. R. Soc. Lond. B* **269,** 227–233 (2002).
4. Dietz, K. Epidemiologic interference of virus populations. *J. Math. Biol.* **8,** 291–300 (1979).
5. Creighton, C. *A History of Epidemics in Britain* (Cambridge Univ. Press, Cambridge, 1894).
6. Rohani, P., Earn, D. J. D., Finkenstadt, B. F. & Grenfell, B. T. Population dynamic interference among childhood diseases. *Proc. R. Soc. Lond. B* **265,** 2033–2041 (1998).
7. Anderson, R. M. & May, R. M. *Infectious Diseases of Humans: Dynamics and Control* (Oxford Univ. Press, Oxford, 1991).
8. Earn, D. J. D., Rohani, P., Bolker, B. M. & Grenfell, B. T. A simple model for complex dynamical transitions in epidemics. *Science* **287,** 667–670 (2000).
9. Rand, D. A. & Wilson, H. B. Chaotic stochasticity: a ubiquitous source of unpredictability in epidemics. *Proc. R. Soc. Lond. B* **246,** 179–184 (1991).
10. McLean, A. & Anderson, R. Measles in developing countries part I. Epidemiological parameters and patterns. *Epidemiol. Infect.* **100,** 111–133 (1988).
11. Anderson, R. M. & May, R. M. Directly transmitted infectious diseases: control by vaccination. *Science* **215,** 1053–1060 (1982).
12. Schenzle, D. An age-structured model of pre- and post-vaccination measles transmission. *IMA J. Math. Appl. Med. Biol.* **1,** 169–191 (1984).
13. Rohani, P., Earn, D. J. D. & Grenfell, B. T. Opposite patterns of synchrony in sympatric disease metapopulations. *Science* **286,** 968–971 (1999).
14. Butler, W. Measles. *Proc. R. Soc. Med.* **6,** 120–153 (1913).
15. Kamo, M. & Sasaki, A. The effects of cross-immunity and seasonal forcing in a multi-strain epidemic model. *Physica D* **165,** 228–241 (2002).
16. Wenjie, W. Control of dengue/dengue haemorrhagic fever in china. *Dengue Bull.* **21** ⟨http://w3.whosea.org/DengueBulletin21/ch3f.htm⟩ (1997).
17. Focks, D. A., Brenner, R. J., Hayes, J. & Daniels, E. Transmission thresholds for dengue in terms of *Aedes aegypti* pupae per person with discussion of their utility in source. *Am. J. Trop. Med. Hyg.* **62,** 11–18 (2000).
18. Hales, S., de Wet, N., Maindonald, J. & Woodward, A. Potential effect of population and climate changes on global distribution of dengue fever: an empirical model. *Lancet* **360,** 830–834 (2002).
19. Kurane, I., Mady, B. J. & Ennis, F. A. Antibody-dependent enhancement of dengue virus infection. *Rev. Med. Virol.* **1,** 211–222 (1991).
20. Behrman, R. E. & Kliegman, R. M. *Nelson Essentials of Pediatrics* (Saunders, Philadelphia, 1998).
21. Cherry, J. D. Pertussis in adults. *Ann. Intern. Med.* **128,** 64–66 (1998).
22. Miller, E. & Gay, N. Epidemiological determinants of pertussis. *Dev. Biol. Stand.* **89,** 15–23 (1997).
23. Keeling, M. J., Rohani, P. & Grenfell, B. T. Seasonally forced disease dynamics explored as switching between attractors. *Physica D* **148,** 317–335 (2001).
24. Butler, W. Whooping cough and measles. *Proc. R. Soc. Med.* **40,** 384–398 (1947).
25. Bartlett, M. S. Measles periodicity and community size. *J. R. Stat. Soc.* **1,** 48–59 (1957).
26. Soper, H. E. The interpretation of periodicity in disease prevalence. *J. R. Stat. Soc.* **92,** 34–73 (1929).
27. Linnert, L. *A statistical report on measles notifications in Manchester, 1917–1951.* (Department of Mathematical Statistics, Manchester, UK, 1954).
28. Grenfell, B. T., Bjornstad, O. N. & Kappey, J. Travelling waves and spatial hierarchies in measles epidemics. *Nature* **414,** 716–723 (2001).
29. Torrence, C. & Compo, G. P. A practical guide to wavelet analysis. *Bull. Am. Meteorol. Soc.* **79,** 61–78 (1998).
30. Buonaccorsi, J. P., Elkington, J. S., Evans, S. R. & Liebhold, A. M. Measuring and testing for spatial synchrony. *Ecology* **82,** 1668–1679 (2001).

Supplementary Information accompanies the paper on *Nature*'s website (**http://www.nature.com/nature**).

Acknowledgements We thank O. Bjornstad, M. Boots, D. Gubler and H. Wearing for comments on this manuscript.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to P.R. (e-mail: rohani@uga.edu).

Unique physiological and pathogenic features of *Leptospira interrogans* revealed by whole-genome sequencing

Shuang-Xi Ren*†‡, Gang Fu*§‡, Xiu-Gao Jiang||‡, Rong Zeng†‡, You-Gang Miao†, Hai Xu†, Yi-Xuan Zhang†, Hui Xiong§, Gang Lu*, Ling-Feng Lu*, Hong-Quan Jiang*§, Jia Jia*, Yue-Feng Tu*, Ju-Xing Jiang¶, Wen-Yi Gu*, Yue-Qing Zhang*#, Zhen Cai*, Hai-Hui Sheng*, Hai-Feng Yin*, Yi Zhang*, Gen-Feng Zhu*, Ma Wan||, Hong-Lei Huang||, Zhen Qian*, Sheng-Yue Wang*, Wei Ma†, Zhi-Jian Yao¶, Yan Shen¶, Bo-Qin Qiang¶, Qi-Chang Xia†, Xiao-Kui Guo§, Antoine Danchin☆, Isabelle Saint Girons, Ronald L. Somerville††, Yu-Mei Wen#, Man-Hua Shi||‡‡, Zhu Chen*§, Jian-Guo Xu|| & Guo-Ping Zhao*†**

** Chinese National Human Genome Center at Shanghai (CHGCS), 250 Bi Bo Road, Zhang Jiang High Tech Park, Shanghai 201203, China*
† Bioinformation Center/Institute of Biochemistry and Cell Biology/Institute of Plant Physiology and Ecology/Research Center of Biotechnology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yue Yang Road, Shanghai 200031, China
§ Rui Jin Hospital/Department of Microbiology and Parasitology, Shanghai Second Medical University, 280 Chongqingnan Road, Shanghai 200025, China
|| National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention (ICDC, China CDC), P.O. Box 5, Changping, Beijing 102206, China
¶ Chinese National Human Genome Center, Beijing, 707 North Yongchang Road, Yi Zhuang High Tech Park, Beijing 100170, China
 # Department of Molecular Virology, Medical Center, Fudan University, 138 Yi Xue Yuan Road, Shanghai 200032, China
☆ HKU-Pasteur Research Centre, 8, Sassoon Road, Hong Kong, China
*** Unité de Bactériologie Moléculaire et Médicale, Institut Pasteur, 25, rue du docteur Roux, 75724 Paris Cedex15, France*
†† Department of Biochemistry, Purdue University, West Lafayette, Indiana 47907, USA
‡ These authors contributed equally to this work
‡‡ Deceased

Leptospirosis is a widely spread disease of global concern. Infection causes flu-like episodes with frequent severe renal and hepatic damage, such as haemorrhage and jaundice. In more severe cases, massive pulmonary haemorrhages, including fatal sudden haemoptysis, can occur[1]. Here we report the complete genomic sequence of a representative virulent serovar type strain (Lai)[2] of *Leptospira interrogans* serogroup Icterohaemorrhagiae consisting of a 4.33-megabase large chromosome and a 359-kilobase small chromosome, with a total of 4,768 predicted genes. In terms of the genetic determinants of physiological characteristics, the facultatively parasitic *L. interrogans* differs

extensively from two other strictly parasitic pathogenic spirochaetes, *Treponema pallidum*[3] and *Borrelia burgdorferi*[4], although similarities exist in the genes that govern their unique morphological features. A comprehensive analysis of the *L. interrogans* genes for chemotaxis/motility and lipopolysaccharide synthesis provides a basis for in-depth studies of virulence and pathogenesis. The discovery of a series of genes possibly related to adhesion, invasion and the haematological changes that characterize leptospirosis has provided clues about how an environmental organism might evolve into an important human pathogen.

Spirochaetes, morphologically unique in their coiled, slender and flexuous shape and related form of motility, form a major phylogenetic lineage (phylum) of eubacteria. *Leptospira*, an obligately aerobic, tightly coiled spirochaete, is the only genus other than *Borrelia*, *Treponema* and *Brachyspira* that is able to cause significant infection in mammals. The leptospires are physiologically chemoheterotrophic. They include the saprophytic *L. biflexa* and the pathogenic *L. interrogans*. The latter is known worldwide to be responsible for the water-borne zoonosis leptospirosis. Although antibiotic therapy is effective against the disease, it remains a serious threat in tropical and subtropical countries as well as in those cities where sanitation is substandard and where wild rats can serve as reservoirs when sewage disposal is poor[1].

Molecular and cellular studies on leptospires[5] have focused on their dynamics of motility, biosynthesis of amino acids and lipopolysaccharide (LPS), outer-membrane proteins and other potential virulence factors. In contrast to *L. biflexa*, little in the way of genetic analysis has been reported for *L. interrogans*, owing to their fastidious cultivation requirements and the lack of genetic systems[5]. Previously, the genomic sequences of two pathogenic spirochaetes—*T. pallidum*, responsible for syphilis[3], and *B. burgdorferi*, responsible for Lyme disease[4]—have been determined. We employed the whole-genome random sequencing method[3,4,6] to sequence and analyse the genomic DNA of a representative virulent serovar type strain (Lai)[2] of *L. interrogans* serogroup Icterohaemorrhagiae (see Methods).

The *L. interrogans* genome (4,691,184 base pairs (bp); Fig. 1, Table 1) is much larger than either of the other two spirochaetes (1,138,006 bp for *T. pallidum* and 1,519,857 bp for *B. burgdorferi*, including plasmids). It consists of two circular chromosomes, a large one of 4,332,241 bp (CI) and a small one of 358,943 bp (CII), in good agreement with previous estimates[5]. More than 30 copies of repetitive DNA elements, including members of the IS1500 and IS1501 families, were distributed throughout the genome but few phage-related sequences were identified.

Both GC nucleotide skew ((G − C)/(G + C)) analysis and comparisons with the *ori* sequences of other bacteria were employed to locate the replication origin of CI, whereas only GC nucleotide skew analysis was used to identify a putative replication origin on CII, as with *Vibrio cholerae*[6]. DnaA boxes were identified on the anticlockwise side of *ori*s for both CI and CII. In addition, *parAB* operons were identified on each side of the putative replication origins of both chromosomes (Supplementary Information 1).

Table 1 **General features of the *L. interrogans* chromosomes**

Features of *L. interrogans* chromosomes	CI	CII	CI + CII
Genome size (bp)	4,332,241	358,943	4,691,184
G + C (%)	36.00	36.10	36.00
Protein coding (%)	78.30	79.80	78.40
Protein-coding genes (no.)	4,360	367	4,727
CDSs with functional assignment	1,901	159	2,060
CDSs without functional assignment	2,459	208	2,667
Unknown functional proteins	140	6	146
Similar to unassigned CDSs predicted in other organisms	509	60	569
No significant similarity to CDSs predicted in other organisms	1,810	142	1,952
Gene density (bp per gene)	993	978	992
Average gene length (bp)	778	781	778
Average unknown gene length (bp)	557	567	558
Insertion sequence (no.)	18	12	30
IS 1500 family	7	1	8
IS 1501 family	1	4	5
Others	10	7	17
Ribosomal RNAs	4	0	4
Transfer RNAs	37	0	37

In all, 4,768 putative genes were predicted, among them 37 genes for transfer RNAs (Supplementary Information 2-1). Previous reports[5] indicated that in strains Ictero No. 1, Verdun and RZ11 of *L. interrogans*, there were two sets each of genes encoding 16S ribosomal RNA (*rrs*) and 23S rRNA (*rrl*). However, besides the two *rrs* genes, we identified only one gene each encoding 5S (*rrf*) and 23S rRNAs respectively. The extraordinarily low number of tRNA and rRNA genes might well account for the fastidious growth of *L. interrogans*.

Among the 4,727 protein-coding sequences (CDSs), 4,360 lie on CI and 367 lie on CII, whereas all of the rRNA and tRNA genes were found on CI (Table 1). Although most of the genes required for growth and viability are located on CI, some essential genes lie on CII. Besides the previously recognized *metF*[7] (LB002) and *asd*[5] (LB355), it is significant to recognize an *ndh* gene (LB036), encoding NADH dehydrogenase, and clusters of genes involved in a nearly complete pathway for the *de novo* biosynthesis of haem. These data, therefore, tend to support the view that CII is an authentic part of the genome that did not originate by lateral transfer.

On the basis of amino acid sequence similarity searches and/or domain analysis, biological functions have been assigned to about 44% of the CDSs (2,060), whereas 15% of the CDSs (715) either encode proteins of unknown function or are similar to unassigned CDSs predicted in other organisms. A total of 1,952 predicted CDSs (41%) failed to exhibit obvious similarity to any protein-coding genes of other organisms (Table 1). In particular, only 315 orthologues were shared by *L. interrogans*, *T. pallidum* and *B. burgdorferi* (Supplementary Information 3).

Some of the previously identified metabolic characteristics of leptospires, such as the absence of hexokinase[1], were confirmed by genomic analysis. A complete set of genes for a system of long-chain fatty-acid utilization, a tricarboxylic acid cycle and a respiratory electron transport chain were identified in *L. interrogans*; this was consistent with the notion that the organism generates ATP by oxidative phosphorylation (Fig. 2). In contrast, none of the aforementioned genes are present in *T. pallidum* or *B. burgdorferi*, in which ATP can be generated only by sugar fermentation by means of the Embden–Meyerhof pathway[5]. Because *L. interrogans* cannot utilize sugars as carbon sources, anaplerotic reactions are essential for gluconeogenesis. We failed to identify genes encoding glucose-6-phosphate dehydrogenase, one of the key enzymes of the phosphogluconate pathway. Neither of the two key enzymes of the glyoxylate pathway, isocitrate lyase and malate synthase, were present, although these two enzymes were detected in *L. biflexa*[1]. However, we did identify all the genes encoding enzymes for gluconeogenesis from glycerol (Fig. 2), including phosphoglucose isomerase, as previously reported[1]. In addition, genes encoding enzymes likely to be involved in the oxidative carboxylation of acetyl-CoA to succinyl-CoA through the 3-hydroxypropionate pathway[8] were recognized (Fig. 2). Intermediates of carbohydrate metabolism are therefore likely to be synthesized by means of the tricarboxylic acid cycle and the non-oxidative pentose phosphate pathway (Fig. 2). Genes encoding transhydrogenase (*pntA* and *pntB*) were identified. These enzymes could catalyse the formation of sufficient NADPH for anabolic processes at the cost of protonmotive force generated by an NADH dehydrogenase complex (Fig. 2). In this connection, one should emphasize that glycerol, together with the long-chain fatty acids, is present in EMJH medium (Johnson and Harris modification of the Ellinghausen and McCullough medium)[1] for better growth of *L. interrogans*.

In contrast to *B. burgdorferi* and *T. pallidum*, *L. interrogans*

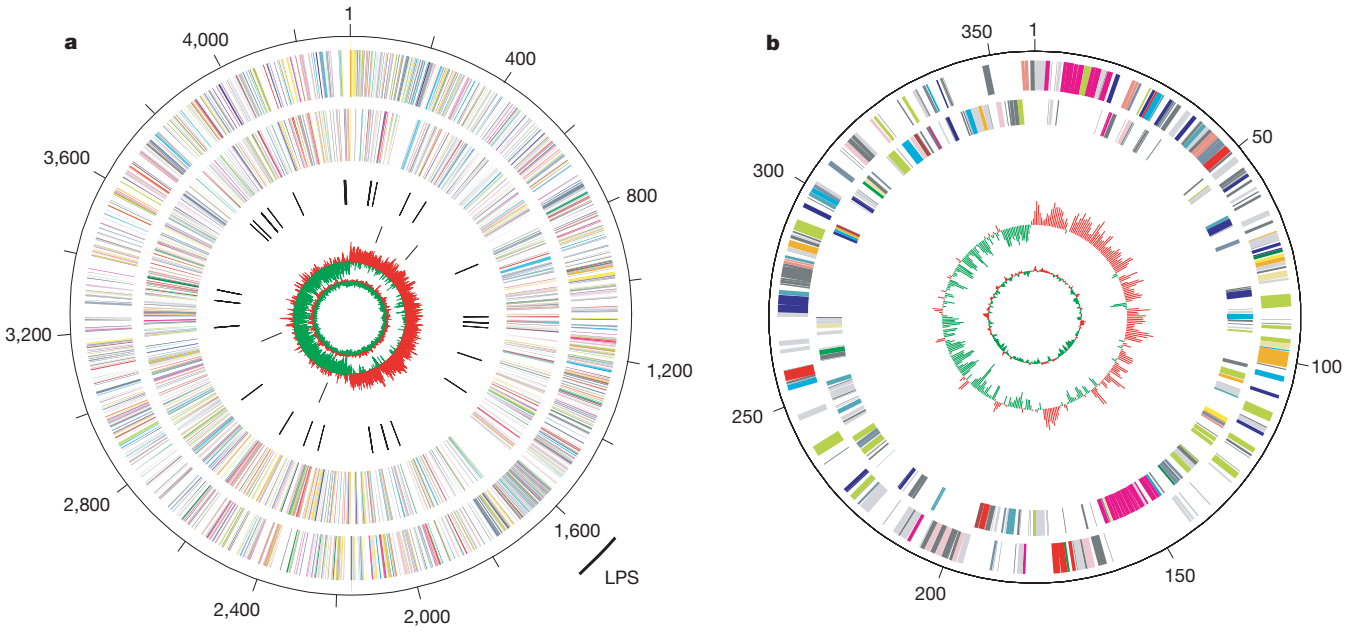

Figure 1 Circular representation of the *L. interrogans* strain Lai genome, with predicted CDSs. **a**, Large chromosome (CI); **b**, small chromosome (CII). The outer scale is shown in kilobases. Circles range from 1 (outer circle) to 6 (inner circle) for CI and from I (outer circle) to IV (inner circle) for CII. Circles 1/I and 2/II, genes on forward and reverse strand; circles 3, tRNA genes; circle 4, rRNA genes; circle 5/III, GC bias ((G-C)/(G + C); red indicates values > 0; green indicates values < 0); circles 6/IV, G + C content. All genes are colour-coded according to functions: orange for amino acid biosynthesis, green for purines, pyrimidines, nucleosides and nucleotides, blue for fatty acid and phospholipid metabolism, magenta for biosynthesis of cofactors, prosthetic groups and carriers, khaki for central intermediary metabolism, cyan for energy metabolism, orchid for transport and binding proteins, yellow for DNA metabolism, dark green for transcription, brown for protein synthesis, red for protein fate, green–yellow for regulatory functions, pink for cell envelope, salmon for cellular processes, navy for other categories, light grey for conserved, dim grey for hypothetical, slate grey for unknown function protein, and black for tRNA and rRNA.

encodes complete metabolic systems for amino acid and nucleotide biosynthesis, which is in agreement with previous work[1]. Methionine biosynthesis in leptospires is similar to that in yeast[1], whereas the final step seems to be catalysed by a B_{12}-dependent homocysteine-N^5-methyltetrahydrofolate transmethylase, encoded by *metH*, rather than by a cobalamin-independent methionine synthase encoded by *metE* (Fig. 2). In this connection, the absence of several genes of B_{12} biosynthesis from the *L. interrogans* genome accounts for the fact that this compound is an essential component of the EMJH semi-synthetic medium[1]. It was proposed that a pyruvate pathway might be used by leptospires for isoleucine biosynthesis, either alone or together with the conventional threonine deaminase pathway[1]. Because we failed to identify a gene encoding threonine deaminase but did find three putative *leuA* genes, we experimentally determined the substrate specificity of these enzymes (see Methods). The enzyme encoded by LA2202 is an isopropylmalate synthase (*leuA*1), whereas LA2350 encodes citramalate synthase (*cimA*). Although the enzyme encoded by LA0469 has some citramalate synthase activity, it is primarily an isopropylmalate synthase (*leuA*2).

The genomic information enhances our understanding of the mechanisms of virulence and pathogenesis in leptospirosis. As with most other pathogenic bacteria, *L. interrogans* possesses several genes related to the attachment and invasion of eukaryotic cells (*mce*, *invA*, *atsE* and *mviN*; Supplementary Information 2-2). The unique cellular shape and motility apparatus of spirochaetes provide these organisms with an additional method of achieving effective infection[5,9]. We found at least 50 genes (not including chemotaxis genes) related to motility, accounting for more than 1% of the deduced CDSs (Fig. 2). Like *B. burgdorferi* and *T. pallidum*, *L. interrogans* uses FlaA sheath protein and FlaB core protein as the essential components of its endoflagellar filament[5]. Other bacteria[5,10] employ FliC for this purpose. *L. interrogans* also has a complete set of genes (Supplementary Information 2-2) for shape determination. In contrast to *B. burgdorferi*[11], the finely coiled spiral shape of leptospires is likely to be mainly attributable to the murein layer rather than the flagella[12].

Chemotaxis is generally acknowledged to be an important virulence factor for pathogenic bacteria. The chemotaxis system of *L. interrogans* (Fig. 2) is more complex than that of either *T. pallidum* or *B. burgdorferi*. The recognition of many genes (12 CDSs) encoding methyl-accepting chemotaxis proteins (MCPs) presumably reflects the extremely diverse environmental situations that a facultatively parasitic zoonotic bacterium can encounter. Employing secondary-structure prediction methods, 5 of the 15 CDSs with clear CheY-like response domains were designated *cheY* genes (Supplementary Information 4-1). However, only one such gene was located in a putative chemotaxis operon (*cheWABY*, LA1250-1253).

Leptospirosis virulence has been attributed in part to the effect of the leptospiral LPS[1]. The nucleotide sequence of the locus encoding a set of enzymes for the biosynthesis of the O-antigen component of *Leptospira* LPS (*rfb* locus) is known for four serovars of two species[5]. We identified an *rfb* locus of 103 kilobases (kb) (Supplementary Information 5) in *L. interrogans* serovar lai. In agreement with findings in other *rfb* loci of leptospires, almost all of the 97 CDSs (LA1576 to LA1672), except three short ones, are encoded on the same strand (forward). About 30 kb of nucleotide sequence located at the 3′-proximal end of the locus is almost identical to its counterpart in serovar copenhageni (GB: U61226). Unlike *L. borgpetersenii* serovar hardjo, subtype hardjobovis, no IS elements were found within or flanking the *rfb* locus. We tentatively assigned a series of genes encoding O-antigen-processing enzymes within and outside the *rfb* locus by comparisons of predicted transmembrane patterns with genes characterized in other Gram-negative bacteria (Supplementary Information 4-2). This is a strong

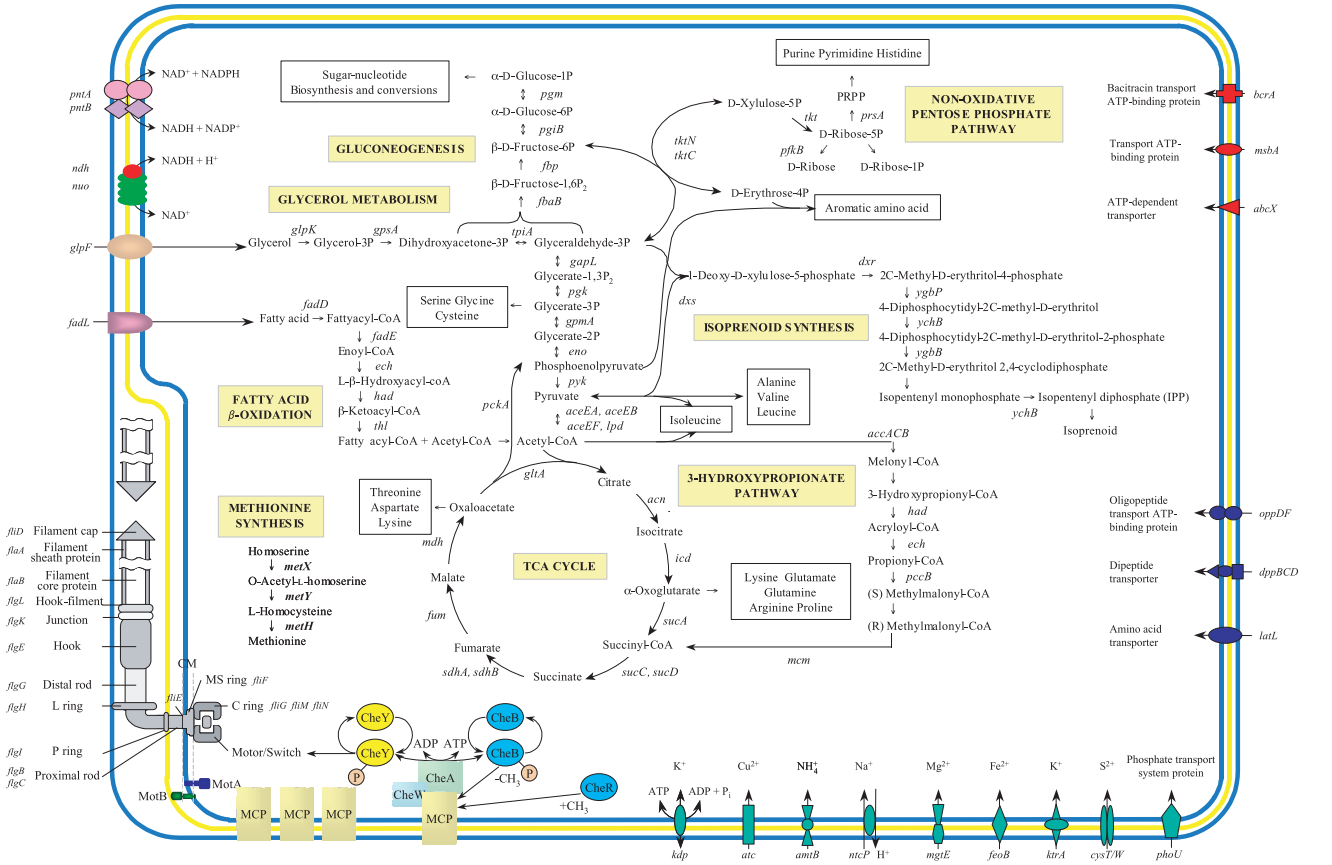

Figure 2 Overview of selected metabolic pathways and morphological components in *L. interrogans* strain Lai. Only pathways related to energy production, biosynthesis of carbon skeletons and certain amino acids (methionine and isoleucine) are shown in detail. In addition, important components of transport systems as well as chemotaxis and motility systems are illustrated. The only two endoflagella are located between the outer membrane sheath (blue) and the cell wall (yellow)/cytoplasmic membrane (blue). At each end of the protoplasmic cell cylinder, a single periplasmic flagellum extends towards the centre of the cell with no overlap between them. Key metabolic enzymes and other related functional proteins are labelled according to their corresponding genes with their CDS numbers listed in Supplementary Information 2-2.

indication that the biosynthesis of LPS in *L. interrogans* proceeds through the Rfc (Wzy)-dependent pathway.

In contrast to *T. pallidum* and *B. burgdorferi*[5], genes encoding enzymes involved in the biosynthesis of the Lipid A backbone and its KDO (2-keto-3-deoxyoctonoic acid) core (Supplementary Information 6) are present in *L. interrogans*. The LPS of *L. interrogans* is a structurally unique molecule of relatively low toxicity[5] that activates macrophages in a distinct manner[13]. These characteristics can be rationalized on the basis of structural comparisons between LpxA proteins of different bacterial origins (Supplementary Information 4-3).

Although it is not clear whether the extensively studied sphingomyelin-specific phospholipases have significant roles in the pathogenesis of leptospirosis[1], we identified four genes encoding other kinds of haemolysin in addition to five genes coding for sphingomyelinase-like proteins (Supplementary Information 7). All these proteins have been expressed in *Escherichia coli*, and their haemolytic activities have been demonstrated (Y.X.-Z. and G.-P.Z., unpublished results).

The genome of *L. interrogans* encodes several proteins bearing homology to animal proteins important in haemostasis (Supplementary Information 8). These include a protein that resembles the mammalian platelet-activating factor (PAF) acetylhydrolase[14] (LA2144, *pafAH*) and another that is similar to von Willebrand factor[15] type A domains (LB054 and LB055, *vwa*). No bacterial genomes have hitherto been shown to encode both of these proteins, although they have been separately identified in several bacterial species (Supplementary Information 8). A third gene relevant to

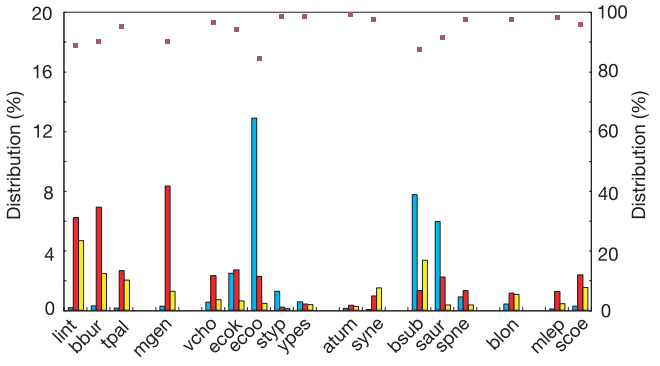

Figure 3 Distribution of the best hits for BLAST protein homologues by representative eubacteria against predicted proteomes of bacteria (brown squares), virus (phage) (blue bars), archaea (yellow bars) and eukarya (red bars). See the Methods section for details of analysis. The symbols used are: lint, *L. interrogans*; bbur, *B. burgdorferi*; tpal, *T. pallidum*; mgen, *M. genitalium*; vcho, *V. cholerae*; ecok, *E. coli* K12; ecoo, *E. coli* O157 Sakai; styp, *Salmonella typhimurium* LT2; ypes, *Yersinia pestis* CO92; atum, *Agrobacterium tumefaciens* C58 Cereon; syne, *Synechocystis* PCC6803; bsub, *B. subtilis*; saur, *Staphylococcus aureus* MW2; spne, *Streptococcus pneumoniae* TIGR4; blon, *Bifidobacterium longum*; mlep, *Mycobacterium leprae*; scoe, *Streptomyces coelicolor*. The percentage distributions of CDSs similar to their counterparts are depicted as coloured histograms. Scales used: 0–100% for bacteria, and 0–20% for virus (phage), archaea and eukarya.

haemostasis, so far found only in *Leptospira*, seems to specify an orthologue of paraoxonase (LA0399, *pon*). This protein might hydrolyse PAF through its arylesterase activity[16]. Because a *colA*[17] gene (LA0872) encoding microbial collagenase has been identified, it is reasonable to propose that collagenase-mediated injury to the vascular epithelium during infection and the subsequent combined effects of the Vwa, PafAH and Pon proteins could lead to a loss of haemostasis, in addition to the proposed effects of LPS[1,13]. This model is consistent with the clinical manifestations of leptospirosis, namely damage to the endothelial cell membranes of small blood vessels[1]. It also might explain the observed sequelae of severe infections by serovar lai, such as massive pulmonary haemorrhage and fatal sudden haemoptysis[1].

Among eubacteria, spirochaetes are evolutionarily primitive[9,18]. However, the fact that leptospires can survive either as saprophytes or as facultative parasites has presumably afforded them significant growth opportunities, although not without pressure for co-evolution in response to their environment or hosts. A BLAST analysis was performed to compare the best-hit distribution of protein homologues in representative eubacteria with the predicted proteomes of bacteria, virus (phage), archaea and eukarya. The result (Fig. 3) suggests that the genome of *L. interrogans* surpasses those of other bacteria in terms of the number of proteins with structural similarity to eukaryal and archaeal proteins that it encodes. In this respect, *L. interrogans* resembles *B. burgdorferi* and *Mycoplasma genitalium*. This raises several important evolutionary questions, including the possibility that lateral gene transfer, operating in parallel with standard gene evolution events, contributed to the emergence of an important human pathogen from an environmental bacterium. □

Methods

Source and culturing of study organism

The *Leptospira interrogans* serogroup Icterohaemorrhagiae serovar lai type strain 56601 used in this study is maintained by the National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention (ICDC, China CDC), Beijing, China[2]. For sequencing purposes, a single colony was picked from EMJH[1] soft agar and cultured in the same medium. The culture thus obtained was then subjected to morphological, serological, genetic and virulence analysis. The properties of the strain were in accordance with those of pathogenic *Leptospira*. For functional analysis, growth curves for *L. interrogans* in EMJH or Korthof[1] medium were measured turbidimetrically, and viable bacterial counts were determined by dark-field microscopy. Culture conditions were then developed to ensure that only mid-exponential-phase bacterial cultures were used for further experimentation.

Genome sequencing and analysis

The genome of strain Lai of *L. interrogans* was sequenced by a whole-genome random sequencing method previously applied to other microbial genomes[3,4,6]. Three different libraries were used in this project. The first two, in pUC18, had inserts of either 1.5–3 kb or 8–10 kb. The third was a 40-kb cosmid library. Altogether, 111,402 sequence reads (Phred value >Q20 (refs 19, 20)) were generated, which gave rise to an overall genome coverage of 8.5 fold, of which 1,600 were from the end sequences of large insert plasmid (8–10-kb) clones and 1,000 were from the end sequences of cosmid clones. The Phred/Phrap/Consed software package[19–21] was used for quality assessment and sequence assembly. The initial assembly yielded 805 contigs, which were clustered into 145 groups based on linking information from forward and reverse sequence reads. Some contigs were also located on the physical map by Southern analysis. Sequence and/or physical gaps of the chromosomes were closed by primer walking and PCR. The final assembly was checked against the physical map of restriction sites, mapped genes and end sequences of large plasmid and cosmid clones.

Assignment of CDSs

CDSs were determined with Glimmer 2.0 (ref. 22) and the Z-curve method[23], and the results were subjected to further manual inspection. A few CDSs were found by hand curating as guided by BLAST results. BLAST searches against the NCBI non-redundant protein database (or SwissProt, PIR and COG) were performed to determine the similarity. The blast search criteria were as follows: (1) e-value $= 10^{-5}$ and (2) at least 60% of the subject sequence was aligned. If there was no database hit, domain analysis was performed by searching the Pfam, PRINTS, PROSITE, ProDom, Block and SMART databases. Transfer RNAs were predicted with tRNAscan-SE[24]. TopPred[25] was used to identify potential membrane-spanning domains in proteins. The presence of signal peptides and the probable position of a cleavage site in secreted proteins were detected with Signal-P. Lipoproteins were identified by scanning for a lipobox ([LV][ASTVI][GAS][C]) in the first 30 amino acids of every protein. Possible metabolic pathways were examined using the KEGG database[10]. Transmembrane helices in proteins were predicted by the THMHH method (Supplementary Information 4). Predicted biological roles were assigned by the classification scheme in ref. 26. In cases in which tertiary structures of hypothetical proteins were predicted, sequences of CDSs were submitted to the SWISS-MODEL server and the illustrations were prepared with Rasmol 2.6.

Deposition of data

In addition to the data deposited at the NCBI database (GB: AE010300 for CI and GB: AE010301 for CII), the *L. interrogans* genome database is also available at http://www.chgc.sh.cn/lep/ and at http://bioinfo.hku.hk/LeptoList/.

BLAST analysis

The BLAST analysis for comparing the best-hit distribution of protein homologues in representative eubacteria with the predicted proteomes of bacteria, virus (phage), archaea and eukarya was based on ref. 27 for studying horizontal gene transfer with modifications. The data were retrieved from NCBI TaxMap (http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/micr.html). The CDSs of each bacterium were used in a BLAST search against the database. Only those hits that scored at least 95 bits were collected and ranked. The 'most' similar organism was the one to which the homologous protein bears the strongest similarity with the query CDS.

Enzyme assays

Citramalate synthase activity was assayed as described in ref. 28, with minor modifications. Isopropylmalate synthase activity was assayed as described in ref. 29, with minor modifications.

Received 23 October 2002; accepted 11 March 2003; doi:10.1038/nature01597.

1. Faine, S., Adler, B., Bolin, C. & Perolat, P. *Leptospira and Leptospirosis* (Medisci, Melbourne, 1999).
2. Kmety, E. & Dikken, H. *Classification of the species Leptospira interrogans and history of its serovars*. (Groningen Univ. Press, The Netherlands, 1993).
3. Fraser, C. M. *et al.* Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* **281,** 375–388 (1998).
4. Fraser, C. M. *et al.* Genomic sequence of a lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* **390,** 580–586 (1997).
5. Saier, M. & Garcia-Lara, J. *The Spirochetes: Molecular and Cellular Biology* (Horizon Scientific Press, Wymondham, 2001).
6. Heidelberg, J. F. *et al.* DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* **406,** 477–483 (2000).
7. Bourhy, P. & Saint Girons, I. Localization of the *Leptospira interrogans metF* gene on the CII secondary chromosome. *FEMS Microbiol. Lett.* **191,** 259–263 (2000).
8. Herter, S. *et al.* Autotrophic CO_2 fixation by *Chloroflexus aurantiacus*: Study of glyoxylate formation and assimilation via the 3-hydroxypropionate cycle. *J. Bacteriol.* **183,** 4305–4316 (2001).
9. Charon, N. W. & Goldstein, S. F. Genetics of motility and chemotaxis of a fascinating group of bacteria: The spirochetes. *Annu. Rev. Genet.* **36,** 47–73 (2002).
10. Ogata, H. *et al.* KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **27,** 29–34 (1999).
11. Motaleb, M. A. *et al. Borrelia burgdorferi* periplasmic flagella have both skeletal and motility functions. *Proc. Natl Acad. Sci. USA* **97,** 10899–10904 (2000).
12. Picardeau, M., Brenot, A. & Saint Girons, I. First evidence for gene replacement in *Leptospira spp.* Inactivation of *L. biflexa flaB* results in non-motile mutants deficient in endoflagella. *Mol. Microbiol.* **40,** 189–199 (2001).
13. Werts, C. *et al.* Leptospiral lipopolysaccharide activates cells through a TLR2-dependent mechanism. *Nature Immunol.* **2,** 346–352 (2001).
14. Arai, H. *et al.* Platelet-activating factor acetylhydrolase (PAF-AH). *J. Biochem. (Tokyo)* **131,** 635–640 (2002).
15. Tuckwell, D. Evolution of von Willebrand factor A (VWA) domains. *Biochem. Soc. Trans.* **27,** 835–840 (1999).
16. Rodrigo, L., Mackness, B., Durrington, P., Hernandez, A. & Mackness, M. Hydrolysis of platelet-activating factor by human serum paraoxonase. *Biochem. J.* **354,** 1–7 (2001).
17. Matsushita, O. *et al.* Gene duplication and multiplicity of collagenases in *Clostridium histolyticum*. *J. Bacteriol.* **181,** 923–933 (1999).
18. Wolf, Y., Rogozin, I., Grishin, N., Tatusov, R. & Koonin, E. V. Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol. Biol.* **1,** 8 (2001).
19. Ewing, B. *et al.* Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8,** 175–185 (1998).
20. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8,** 186–194 (1998).
21. Gordon, D., Abajian, C. & Green, P. Consed: A graphical tool for sequence finishing. *Genome Res.* **8,** 195–202 (1998).
22. Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27,** 4636–4641 (1999).
23. Zhang, C. T. & Wang, J. Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve. *Nucleic Acids Res.* **28,** 2804–2814 (2000).
24. Lowe, T. & Eddy, S. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25,** 955–964 (1997).
25. Claros, M. & von Heijne, G. TopPred II: an improved software for membrane protein structure predictions. *Comput. Appl. Biosci.* **10,** 685–686 (1994).
26. Riley, M. Functions of the gene products of *Escherichia coli*. *Microbiol. Rev.* **57,** 862–952 (1993).
27. Olendzenski, L., Liu, L., Zhaxybayeva, O., Murphey, R., Shin, D. G. & Gogarten, J. P. Horizontal transfer of archaeal genes into the *Deinococcaceae*: Detection by molecular and computer-based approaches. *J. Mol. Evol.* **51,** 587–599 (2000).
28. Howell, D. M., Xu, H. & White, R. H. (R)-Citramalate synthase in methanogenic *Archaea*. *J. Bacteriol.* **181,** 331–333 (1999).

29. Kohlhaw, G. B., Leary, T. R. & Umbarger, H. E. Alpha-isopropylmalate synthase from *Salmonella typhimurium*. Purification and properties. *J. Biol. Chem.* **244,** 2218–2225 (1969).

Supplementary Information accompanies the paper on *Nature*'s website (http://www.nature.com/nature).

Acknowledgements We thank L. Bao, B.-M. Dai, J. Yan, C. Werts, M. Picardeau and G. Baranton for suggestions and comments on our research strategy and manuscript preparation; C. Jin and G.-C. Liu of the Institute of Microbiology, Chinese Academy of Science, for help in the attempt at assaying the enzymatic activity of PafAH; Y. Liu and H.-G. Zhu for help in preparing the drawings; X. Mao and G. Cai for help in computer simulation; B.-Y. Hu and Y-X. Nie for help in bacterial culture preparation; and the members of CHGCS for support and encouragement. This work was supported by the National Natural Science Foundation of China, the Chinese National High Technology Development Program (863), the National Key Program for Basic Research (973) and the Sciences and Technology Commission of the People's Government of Shanghai Municipality. It was also supported by the Pôle Sino-Français en Sciences du Vivant et en Génomique and le Programme de Recherches Avancées Franco-Chinois PRA B00-05.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.-P.Z. (e-mail: gpzhao@sibs.ac.cn). The sequences have been submitted to NCBI under accession numbers GB: AE010300 and GB: AE010301 for the large (CI) and small (CII) chromosomes, respectively.

The methylated component of the *Neurospora crassa* genome

Eric U. Selker*, Nikolaos A. Tountas†‡, Sally H. Cross†‡, Brian S. Margolin*‡, Jonathan G. Murphy*, Adrian P. Bird† & Michael Freitag*

* *Department of Biology and Institute of Molecular Biology, University of Oregon, Eugene, Oregon 97403, USA*
† *Institute of Cell and Molecular Biology, University of Edinburgh, Edinburgh, EH9 3JR, UK*

Cytosine methylation is common, but not ubiquitous, in eukaryotes. Mammals[1] and the fungus *Neurospora crassa*[2,3] have about 2–3% of cytosines methylated. In mammals, methylation is almost exclusively in the under-represented CpG dinucleotides, and most CpGs are methylated[1] whereas in *Neurospora*, methylation is not preferentially in CpG dinucleotides and the bulk of the genome is unmethylated[4]. DNA methylation is essential in mammals[5] but is dispensable in *Neurospora*[3,6], making this simple eukaryote a favoured organism in which to study methylation. Recent studies indicate that DNA methylation in *Neurospora* depends on one DNA methyltransferase, DIM-2 (ref. 6), directed by a histone H3 methyltransferase, DIM-5 (ref. 7), but little is known about its cellular and evolutionary functions. As only four methylated sequences have been reported previously in *N. crassa*, we used methyl-binding-domain agarose chromatography[8] to isolate the methylated component of the genome. DNA sequence analysis shows that the methylated component of the genome consists almost exclusively of relics of transposons that were subject to repeat-induced point mutation—a genome defence system that mutates duplicated sequences[9].

To isolate the methylated component of the *N. crassa* genome, we cleaved genomic DNA with the 5-methylcytosine-sensitive restriction enzyme *Sau*3AI (recognition sequence GATC) so as to leave intact patches of methylated DNA, and then passed it over a methyl-CpG domain (MBD) column, which fractionates according to the degree of CpG methylation[8]. Bound DNA was eluted with increasing concentrations of salt, and fractions were analysed by Southern hybridizations, probing for an unmethylated sequence (*am*) and previously identified methylated regions ($\zeta - \eta$, Ψ_{63} and ribosomal DNA; Fig. 1). DNA complementary to the *am* probe eluted principally in pool four but trailed through to pool nine. In contrast, $\zeta - \eta$ sequences peaked later, in pool nine, suggesting that the MBD column successfully fractionated *Neurospora* DNA on the basis of methylation. Considering that the MBD does not bind methylated non-CpG sites[10], these findings suggest co-localization of methylated CpG and non-CpG (*Sau*3AI) sites. By this assay, the

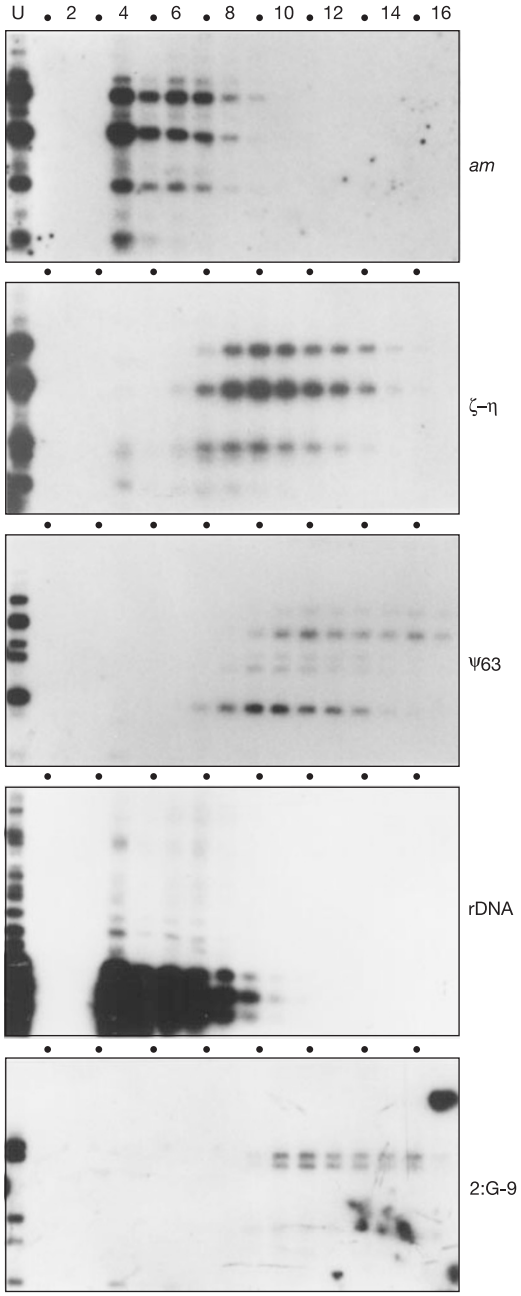

Figure 1 Fractionation of *Neurospora* DNA on a methylated-DNA-binding column. Samples (about 0.5 μg) of pooled pairs of fractions off the MBD column (1–16 represent column fractions 1–32) were fractionated by agarose gel electrophoresis, along with an unfractionated (U) sample of the *Sau*3AI-digested DNA, blotted to nylon membrane, and probed sequentially for known unmethylated (*am*) and methylated ($\zeta - \eta$, Ψ_{63} and rDNA) sequences, as well as a candidate methylated sequence from this study (2:G-9).

‡ Present addresses: Center for Cell Signaling, University of Virginia School of Medicine, Charlottesville, Virginia 22908, USA (N.A.T.); MRC Human Genetics Unit, Western General Hospital, Edinburgh, EH4 2XU, UK (S.H.C.); Department of Biochemistry and Biophysics, University of California San Francisco, California 94143, USA (B.S.M.).