Complete Genome Sequence of *Neisseria meningitidis* Serogroup B Strain MC58

Hervé Tettelin,[1*] Nigel J. Saunders,[2] John Heidelberg,[1]
Alex C. Jeffries,[2] Karen E. Nelson,[1] Jonathan A. Eisen,[1]
Karen A. Ketchum,[1†] Derek W. Hood,[2] John F. Peden,[2]
Robert J. Dodson,[1] William C. Nelson,[1] Michelle L. Gwinn,[1]
Robert DeBoy,[1] Jeremy D. Peterson,[1] Erin K. Hickey,[1]
Daniel H. Haft,[1] Steven L. Salzberg,[1] Owen White,[1]
Robert D. Fleischmann,[1] Brian A. Dougherty,[1] Tanya Mason,[1]
Anne Ciecko,[1] Debbie S. Parksey,[1] Eric Blair,[1] Henry Cittone,[1]
Emily B. Clark,[1] Matthew D. Cotton,[1] Terry R. Utterback,[1]
Hoda Khouri,[1] Haiying Qin,[1] Jessica Vamathevan,[1] John Gill,[1]
Vincenzo Scarlato,[3] Vega Masignani,[3] Mariagrazia Pizza,[3]
Guido Grandi,[3] Li Sun,[2] Hamilton O. Smith,[1†] Claire M. Fraser,[1]
E. Richard Moxon,[2] Rino Rappuoli,[3] J. Craig Venter[1†]

The 2,272,351–base pair genome of *Neisseria meningitidis* strain MC58 (serogroup B), a causative agent of meningitis and septicemia, contains 2158 predicted coding regions, 1158 (53.7%) of which were assigned a biological role. Three major islands of horizontal DNA transfer were identified; two of these contain genes encoding proteins involved in pathogenicity, and the third island contains coding sequences only for hypothetical proteins. Insights into the commensal and virulence behavior of *N. meningitidis* can be gleaned from the genome, in which sequences for structural proteins of the pilus are clustered and several coding regions unique to serogroup B capsular polysaccharide synthesis can be identified. Finally, *N. meningitidis* contains more genes that undergo phase variation than any pathogen studied to date, a mechanism that controls their expression and contributes to the evasion of the host immune system.

Neisseria meningitidis (meningococcus), a Gram-negative β-Proteobacterium (a class that includes *Bordetella*, *Burkholderia*, *Kingella*, and *Methylomonas*), is a cause of life-threatening invasive bacterial infections, especially in young infants. The major diseases caused by *N. meningitidis*, meningitis and septicemia, are a significant public health problem and are responsible for deaths and disability through epidemics in sub-Saharan Africa, and sporadic cases that are prevalent in many countries worldwide (*1*). There are five pathogenic *N. meningitidis* serogroups (A, B, C, Y, and W135) as determined by capsular polysaccharide typing (*2*). Some disease control has been achieved through vaccination, but its impact has been limited. The pattern of serogroup B disease is typically

hyperendemic or sporadic and contrasts with the classically epidemic nature of the serogroup A disease. Strains of serogroup B are a particular problem because they are a major cause of invasive disease in Europe and the United States (*3*), and there is currently no effective vaccine. Sequencing the genome of strain MC58 [a serogroup B strain isolated from a case of invasive infection (*4*)] provides an efficient means of acquiring data relevant to the detailed molecular characterization of this pathogen. Preliminary comparison of its genome sequence to that of the unannotated genome sequence of the serogroup A strain Z2491 (*5*), as well as comparison of the gene complement of strain MC58 with that of *Haemophilus influenzae* (*6*), another pathogen responsible for meningitis and the first for which complete genome sequence data were available, also provides an opportunity to define a common subset of genes that may be responsible for the pathogenesis of this disease.

The complete genome sequence (GenBank accession number AE002098) was obtained by the random shotgun sequencing strategy (*7*). *N. meningitidis* strain MC58 has a genome size of 2,272,351 base pairs (bp) with an average G+C content of 51.5%. Base

pair 1 of the chromosome was assigned within the putative origin of replication that was determined by the presence of a cluster of DnaA boxes, oligomer-skew (*8*), and G-C skew (*9*) analyses. The approach to genome annotation was novel in that it combined the results of open reading frame (ORF) prediction with whole-genome homology searches (*10*). In addition, experiments on a subset of ORFs showed that the products of 85 of these were potentially located on the surface of the meningococcus (*11*).

The genome contains four ribosomal RNA (rRNA) operons (16*S*-23*S*-5*S*) and 59 tRNAs with specificity for all 20 amino acids. The 2158 ORFs identified [Fig. 1; Web figure 1 and table 1 (*12*)] represent 83% of the genome, with an average size of 874 bp. Biological roles were assigned to 1158 ORFs (53.7%) with similarity to proteins of known function according to the classification scheme adapted from Riley (*13*). Three hundred and forty-five (16.0%) predicted coding sequences matched gene products of unknown function from other species, and 532 (24.7%) had no database match (www.tigr.org/tdb/mdb/mdb.html).

We have identified 234 families of proteins in strain MC58 (*14*), containing a total of 678 proteins (32% of the total number of ORFs). The largest family consists of 24 adenosine 5′-triphosphate (ATP)–binding subunits of ABC transporters. The extent of potential recent gene duplications in strain MC58 was estimated by identification of ORFs that are more similar to other ORFs within the strain MC58 genome than to ORFs from other genomes (*15*). This analysis revealed that out of 678 proteins present in families, 105 may have evolved through a process that involved duplication. Most of these are predicted to be involved in pathogenicity functions or transposition activities, or are of unknown function.

Strain MC58 contains 22 intact and 29 remnant insertion sequences (ISs). Most of the latter have transposase genes that contain multiple frame shifts, large deletions and/or premature termination codons, or lack associated inverted repeats. The IS families represented are IS3, IS5, IS30, and IS110 (*16*), and two groups belong to as yet unclassified families; one of these is related to IS1016 (*17*), and the second was first identified in *Synechocystis* sp. (*18*). The majority of putative functional ISs are closely related to IS4351 (*19*) (IS30 family), whereas the other putative functional ISs are related to IS1106A3 (*20*) (IS5 family). Three homologs of the pilin-associated invertases of *N. gonorrhoeae* (*21*) and *Moraxella lacunata* (*22*), which are closely related to transposases, are present. Two of these genes, *pivNM-1A* and *pivNM-1B*, are identical to one another and are similar to the *N. gonorrhoeae* sequence

[1]The Institute for Genomic Research (TIGR), 9712 Medical Center Drive, Rockville, MD 20850, USA. [2]Institute of Molecular Medicine, University of Oxford, Headington, Oxford OX3 9DS, UK. [3]Immunological Research Institute of Siena (IRIS), Chiron S.p.A., Via Fiorentina 1, 53100 Siena, Italy.

*To whom correspondence should be addressed. E-mail: tettelin@tigr.org
†Present address: Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA.

(42.4% identity), whereas the remaining gene, *pivNM-2*, is slightly more divergent (38.8% identity).

Neisserial DNA uptake signal sequences (USSs) for DNA transformation (*23*) play a role in recognition of homospecific DNA during transformation. In strain MC58, a total of 1910 USSs (5'-GCCGTCTGAA-3') distributed throughout the genome have been identified, with an approximately equal number on each strand. Only four USSs are expected on the basis of genome sequence length and base composition. Forty-six percent of the USSs are found as inverted pairs located 3' of ORFs, which could form stem-

loop structures and function as transcriptional terminators.

Molecular phylogenetic studies have indicated that the *Neisseria* genus belongs to the β subgroup of the Proteobacteria. The β-Proteobacteria are closely related to the γ-Proteobacteria, which include *Vibrio*, *Haemophilus*, and *Escherichia*. The availability of complete genome sequences from several proteobacterial species (*H. influenzae*, *Helicobacter pylori*, *Escherichia coli*, *Rickettsia prowazekii*, and *Campylobacter jejuni*) allows an evaluation of the position of *Neisseria* within the Proteobacteria and a better understanding of recent evolutionary

events in the *Neisseria* lineage. Seventy-one percent of the ORFs in strain MC58 are most similar (*24*) to ORFs from γ-Proteobacteria (*E. coli* or *H. influenzae*), indicating that most of the genome is proteobacterial in nature and supporting the hypothesis that the β and γ subgroups share a recent common ancestor. Since the divergence of the β and γ subgroups, there must have been extensive gene loss in the *H. influenzae* lineage and/or gene addition and duplication in the *E. coli* lineage that has led to the much larger genome size in *E. coli*. Relative to the total number of genes in each genome, *N. meningitidis* is more
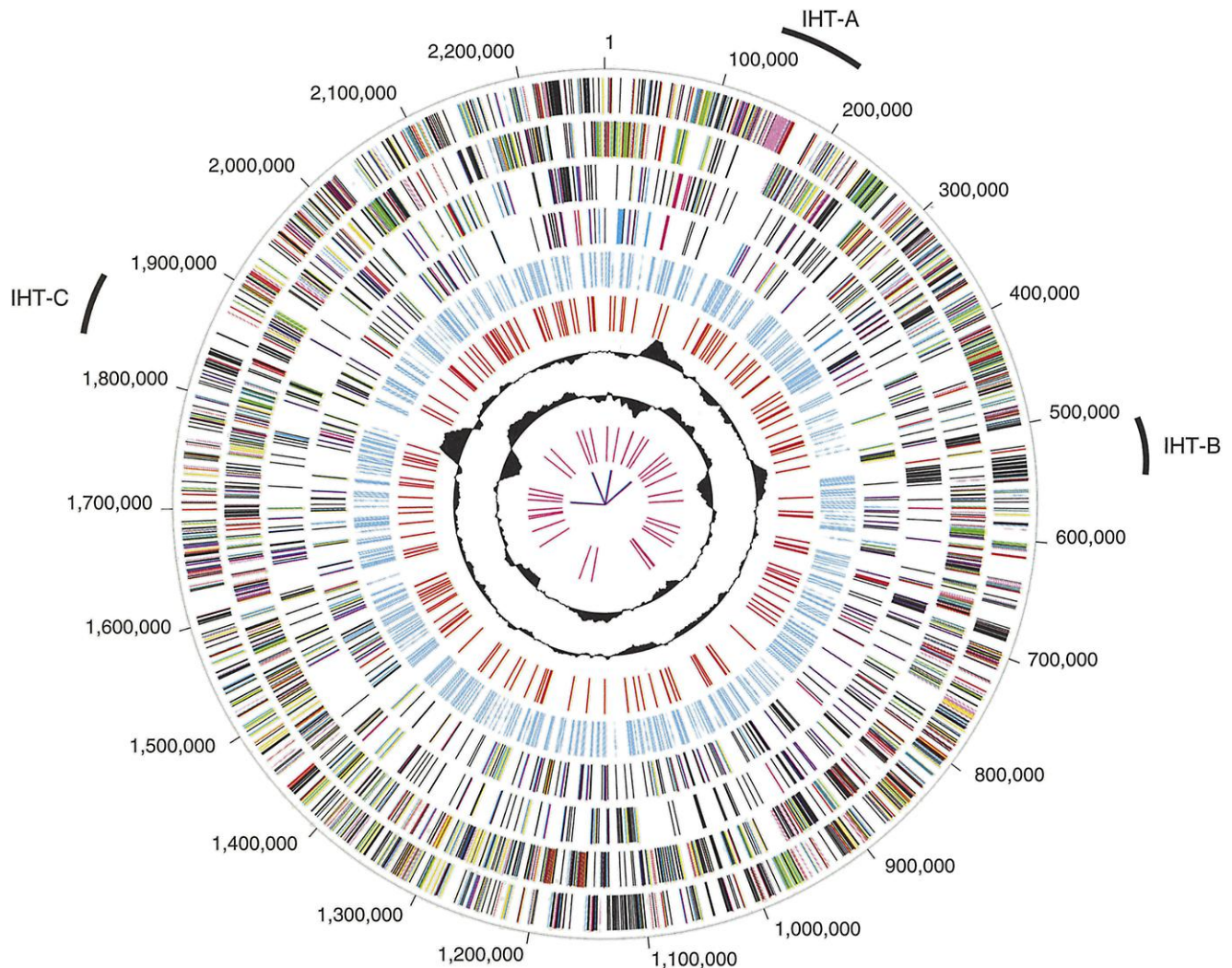

Fig. 1. Circular representation of the *N. meningitidis* strain MC58 genome. Outer circle, predicted coding regions on the plus strand color-coded by role categories [Web figure 1 and table 1 (*12*)]: salmon, amino acid biosynthesis; light blue, biosynthesis of cofactors, prosthetic groups, and carriers; light green, cell envelope; red, cellular processes; brown, central intermediary metabolism; yellow, DNA metabolism; green, energy metabolism; purple, fatty acid and phospholipid metabolism; pink, protein fate and synthesis; orange, purines, pyrimidines, nucleosides, and nucleotides; blue, regulatory functions; gray, transcription; teal, transport and binding proteins; black, hypothetical and conserved hypothetical proteins. Second circle, predicted coding regions on the minus strand color-coded by role categories. Third circle, predicted coding regions

on the plus strand color-coded by function in virulence (Tables 1 and 2): red, acquisition; cyan, colonization; magenta, evasion; yellow, toxins; green, unknown; dark green, putative phase-variable genes; blue, candidate vaccine genes that were expressed in vitro and whose products are potentially located on the surface of the meningococcus (*11*); black, hypothetical and conserved hypothetical proteins. Fourth circle, predicted coding regions on the minus strand color-coded by function in virulence. Fifth circle, *Neisseria* USSs. Sixth circle, *N. gonorrhoeae* inverted repeats (*33*). Seventh circle, atypical nucleotide composition curve [the dinucleotide signatures analysis (*29*) is shown]. Eighth circle, percent G+C curve. Ninth circle, tRNAs. Tenth circle, ribosomal RNAs.

similar to *H. influenzae* than to *E. coli* (Fig. 2). This could be due either to parallel loss of genes in *N. meningitidis* and *H. influenzae* and/or exchanges of genes between these lineages.

Of the 2158 ORFs in serogroup B strain MC58, 1968 (91.2%) are similar (BLASTN $P < 10^{-10}$) to ORFs in the serogroup A strain Z2491 (5) and are likely orthologs. Most of the 190 ORFs without similarity are hypothetical proteins. Comparison between the serogroup B strain MC58 and the serogroup A strain Z2491 genome sequences reveals a major inversion of 955 kb. The scale of this inversion is substantially greater than has been previously described in nonenteric bacteria (25) and also differs from the homologous recombination associated with rRNA operons in enteric bacteria (26). The other main differences between these two strains reside in the regions described below.

Lateral transfer of DNA between species is well documented and is often associated with the evolution of pathogenicity (27). Regions of DNA that have been obtained by lateral gene transfer are often characterized by atypical DNA composition relative to the rest of a genome (28). Therefore, percent G+C, dinucleotide signatures, and χ^2 analyses (29) were used to identify such regions. Three major regions of atypical nucleotide composition were identified in the strain MC58 genome; these have been designated as putative islands of horizontally transferred DNA (IHTs) (Figs. 1 and 3). IHT-A consists of two subregions: IHT-A1 (NMB0066 to NMB0074) contains the genes of the serogroup B capsulation cluster and an adenine rRNA methylase. IHT-A2 (NMB0091 to NMB0100) contains two disrupted ORFs with similarity to an ABC transporter and a secreted protein, and eight hypothetical proteins and is flanked by two disrupted copies of IS1016 (17). IHT-B contains 24 hypothetical proteins (NMB0498 to NMB0521). IHT-C contains 30 ORFs (NMB1746 to NMB1775), several of which encode proteins that may have a role in virulence. These include genes encoding three toxin/toxin-related homologs; a protein known to be immunogenic (30); one intact and three fragmented proteins previously associated with bacteriophage (31); a protein similar to a virulence-associated protein from *H. pylori* (32); two different, apparently intact transposases that do not flank the region; and 19 hypothetical proteins. The transposases do not form a composite transposon based on their location. One transposase is similar to that of IS4351 (19) (of which there are 14 copies in the genome), and the other is *pivNM-2*. IHT-B and IHT-C are devoid of neisserial USSs, and all three IHTs lack an
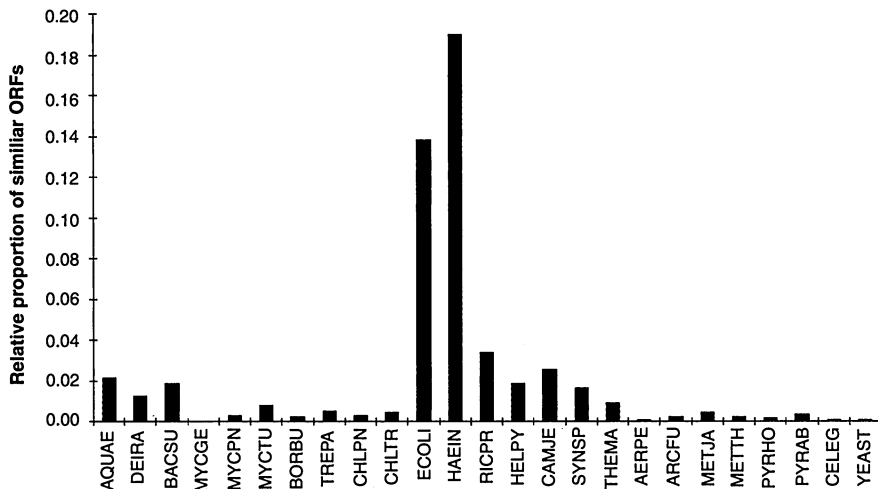

Fig. 2. Comparison of the *N. meningitidis* strain MC58 ORFs to that of other completely sequenced organisms. The sequences of all proteins from each completely sequenced genome were retrieved from the National Center for Biotechnology Information, TIGR, and the *Caenorhabditis elegans* (wormpep16) databases. All *N. meningitidis* ORFs were searched against the ORFs from all other genomes with FASTA3 (24). The number of *N. meningitidis* ORFs whose highest similarity ($p < 10^{-5}$) is to an ORF from a given species is shown in proportion to the total number of ORFs in that species. Abbreviations: AQUAE, *Aquifex aeolicus*; DEIRA, *Deinococcus radiodurans*; BACSU, *Bacillus subtilis*; MYCGE, *Mycoplasma genitalium*; MYCPN, *Mycoplasma pneumoniae*; MYCTU, *Mycobacterium tuberculosis*; BORBU, *Borrelia burgdoferi*; TREPA, *Treponema pallidum*; CHLPN, *Chlamydia pneumoniae*; CHLTR, *Chlamydia trachomatis*; ECOLI, *Escherichia coli*; HAEIN, *Haemophilus influenzae*; RICPR, *Rickettsia prowazekii*; HELPY, *Helicobacter pylori*; CAMJE, *Campylobacter jejuni*; SYNSP, *Synechocystis* sp.; THEMA, *Thermotoga maritima*; AERPE, *Aeropyrum pernix*; ARCFU, *Archaeoglobus fulgidus*; METJA, *Methanococcus jannaschii*; METTH, *Methanobacterium thermoautotrophicum*; PYRHO, *Pyrococcus horikoshii*; PYRAB, *Pyrococcus abyssii*; CELEG, *Caenorhabditis elegans*; YEAST, *Saccharomyces cerevisiae*.
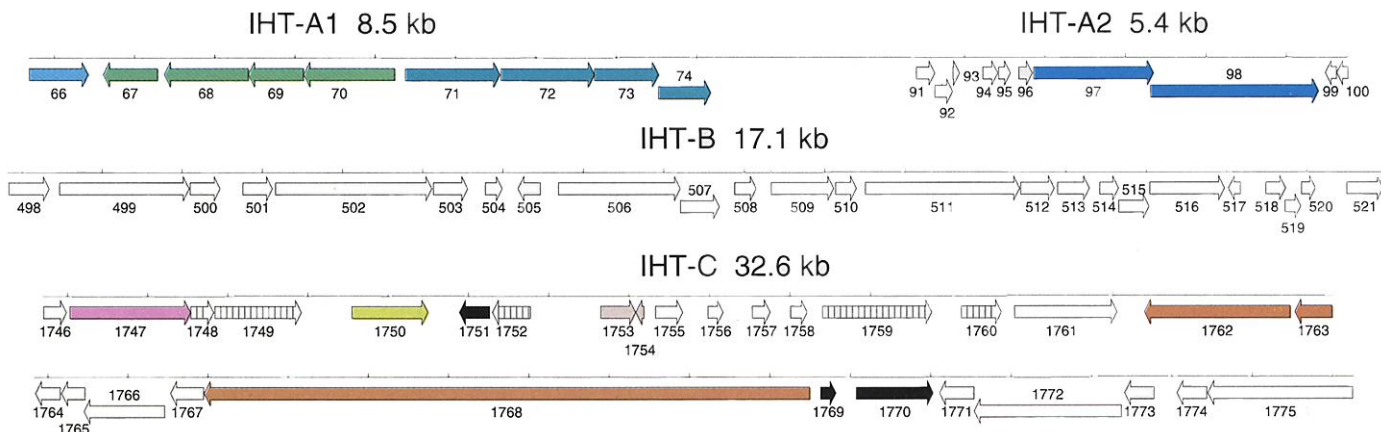

Fig. 3. Structure of the putative islands of horizontally tranferred DNA (IHTs) in the *N. meningitidis* strain MC58 genome. Empty boxes are hypothetical proteins and striped boxes are conserved hypothetical proteins. IHT-A1: NMB0066, adenine rRNA methylase ErmC; NMB0067 to NMB0070, capsule biosynthesis proteins SiaD, SiaC,SiaB, and SynX; NMB0071 to NMB0074, capsule export proteins CtrA, CtrB, CtrC, and CtrD. IHT-A2: NMB0097 and NMB0098, disrupted secreted protein and ABC transporter. IHT-C: NMB1747, tspB protein; NMB1750, PivNM-2; NMB1751, NMB1769, and NMB1770, transposases; NMB1753 and NMB1754, bacteriophage-related proteins; NMB1762, NMB1763, and NMB1768, toxin/toxin-related homologs.

inverted repeat identified in *N. gonorrhoeae* (*33*) that is present in 163 copies in strain MC58, supporting their foreign origin (Fig. 1). The three IHTs in strain MC58 do not have the classical characteristics of "pathogenicity islands" (*34*) since transposase genes flank only IHT-A2, other regions are not flanked by inverted repeats, and none is adjacent to tRNA genes.

Differences in IHTs between *N. meningitidis* serogroup A and serogroup B strains could suggest a different complement of putative virulence determinants. A comparison to the genome sequence of the serogroup A strain Z2491 (*5*) reveals that it contains only one IHT and that it is not present in the strain MC58 genome. Conversely, none of the strain MC58 IHTs are present in the strain Z2491 genome. In contrast with IHT-A1, the genes of the serogroup A capsulation cluster do not generate the same IHT signature, suggesting that the different component genes have been in *Neisseria* for a longer evolutionary time.

A crucial factor in the commensal and pathogenic behavior of *N. meningitidis* is its capacity to obtain and synthesize nutrients essential for its survival [Web figure 2 (*12*)]. Genome analysis supports the observation that maltose and glucose are the only two sugars utilized for energy. The uptake and initial degradation of maltose may be accomplished by a system similar to that described for *Lactobacillus sanfranciscensis* (*35*). In this bacterium, import of maltose is mediated by a maltose/H^+ symport, rather than the ABC transporter (MalEFGK) or PEP:PTS system (MalX) characterized in *E. coli* and many other bacteria. Metabolism of maltose to glucose proceeds by a maltose phosphorylase (NMB0390) distinct from other α-glucosidases. This gene is associated with sequences that encode proteins involved in both sugar metabolism (NMB0389 and NMB0391) and sugar transport (NMB0388), suggesting coordinate regulation of maltose catabolism with substrate availability.

Degradation of glucose, the amino acids serine, proline, and glycine, and the organic acids acetate, gluconate, glutamate, lactate, malate, oxaloacetate, and pyruvate is accomplished by way of an intact tricarboxylic acid (TCA) cycle, previously identified in biochemical studies of *N. meningitidis* (*36*), as well as the pentose phosphate and Entner-Doudoroff pathways. *N. gonorrhoeae* demonstrates metabolic diversity under different growth conditions. This bacterium alters glucose metabolism in response to changes of environmental pH (*37*), with glucose being channeled through glycolysis, the Entner-Doudoroff pathway, or the pentose phosphate pathway, depending on pH fluctuations in the growth medium.

The binding of iron by a number of high-affinity binding proteins in the human body constrains its availability (*38*), and successful pathogens have evolved mechanisms to acquire iron from the host (*39*). Genome analysis reveals that *N. meningitidis* has a large number of systems for scavenging iron, including previously recognized hemoglobin, transferrin, and lactoferrin binding proteins (*38*), as well as additional systems for iron acquisition, including siderophore acceptor and utilization homologs whose functions have not yet been investigated.

Although both *H. influenzae* and *N. meningitidis* reside in the nasopharynx and can cause meningitis, there are metabolic differences between these organisms identified in previous biochemical characterizations (*40*) and extended by genome analysis. *H. influenzae* lacks an intact TCA cycle and the Entner-Doudoroff pathway, has far fewer systems available for the import of iron, and has a smaller percentage of genes devoted to electron transport systems than *N. meningitidis*. *H. influenzae* does, however, have a larger number of transporters devoted to the import of amino acids and carbohydrates including glucose, xylose, ribose, fucose, mannonate, and galactose, suggesting that *H. influenzae* derives more energy through substrate-level phosphorylation than through elaborate electron-transfer systems. The significance of the metabolic differences between *N. meningitidis* and *H. influenzae* is not immediately apparent but may, in part, be responsible for the differential abilities of these two organisms to cause disease under varying physiological conditions in the human host.

When the meningococcus causes invasive disease in a susceptible individual, the process involves invasion of the respiratory tract epithelia and the underlying endothelia of the microvascular system, followed by systemic dissemination through the bloodstream. Efficient replication of meningococci and the elaboration of cell wall and other microbial molecules, such as lipopolysaccharide (LPS) or peptidoglycan, excite inflammation in the host tissues. The outcome is life-threatening septicaemia and metastatic spread to the meninges and cerebrospinal fluid (meningitis).

The genes in *N. meningitidis* known to be related to pathogenicity were classified into several major functional categories (Table 1). Genes encoding proteins involved in colonization of the human respiratory tract epithelium are present in strain MC58. These include genes for the type IV pilus, Opa proteins (four genes), and Opc. Pili are crucial to the niche specialization of the meningococcus because they mediate tropism specifically for human respiratory tract epithelia (*41*). Antigenic variation of pilus proteins is important to this function because it generates the polymorphisms that facilitate the capacity to colonize different individuals and niches within each host and to evade clearance mechanisms (*42*). In both the serogroup B strain MC58 and the serogroup A strain Z2491, the transcriptionally silent partial pilin-coding sequences (eight cassettes) and the expressed gene encoding the complete pilin protein are grouped together in a single locus. Recombination within these cassettes, by a process probably involving associated DNA repeat sequences, generates the antigenic variation of the polymerized pilin proteins. The clustered organization of sequences in *N. meningitidis* contrasts with that found in the closely related pathogen *N. gonorrhoeae*, in which the silent and expressed sequences are distributed throughout the genome (*43*). In the meningococcus, there is a single Sma/Cla repeat downstream of this locus, similar to that documented in *N. gonorrhoeae* (*44*), but the intergenic regions of the meningococcus differ in the number and distribution of RS1, RS2, and RS3 repeats (*43*). In particular, the intergenic region contains a relatively large number of RS3 repeats arranged as inverted pairs. These findings suggest that the *N. meningitidis* and *N. gonorrhoeae* use different mechanisms and sequence substrates for generating antigenic variation of the pilus.

In silico analysis of the genome sequence facilitates the identification of genes involved in surface interactions and virulence, based on specific searches for homologs of those that have been previously characterized in other pathogenic species. These include three large related proteins that contain homologies to adhesins, as well as toxin motifs and an additional type IV pilin protein (NMB0547) that is unrelated to the *pilE/pilS* system. This protein is distinct from the neisserial type II pilus protein previously recognized as an alternative pilus system in *N. meningitidis* (*45*).

The capsular polysaccharide and LPS that impede clearance and killing of meningococci by phagocytes are critical in the pathogenesis of invasive infection and have been previously characterized (*46*). In addition to the genes required for synthesis of the polysialic acid capsule of strain MC58 present in IHT-A1, there are differences compared to the serogroup A sequence in the adjacent capsule exporter genes (*47*). Whereas CtrA and CtrD share >95% identity in their deduced amino acids to their serogroup A homologs, the *ctrB* and *ctrC* genes, organized in an operon between *ctrA* and *ctrD*, encode proteins that have <85% amino acid identity with their homologs in the serogroup A genome. CtrB and CtrC are thought to be membrane-associated components of the capsule export apparatus (*47*). Such divergence may reflect specificity for the serogroup B (polysialic acid) or serogroup A (poly-2-acetamido-2-deoxy-D-mannopyranosyl phosphate) capsular polysaccharides.

Table 1. Putative pathogenicity genes in *N. meningitidis* strain MC58, listed by functional category. Systematic gene names start with NMB (*N. meningitidis* serogroup B), followed by a number indicating the gene position with regard to the origin of replication.

Previously known genes		New genes indentified in strain MC58	
		Acquisition	
NMB0460	Transferrin-binding protein 2 (tbp2)	NMB0293	TonB-dependent receptor, putative
NMB0461	Transferrin-binding protein 1 (tbp1)	NMB0718	Ferrochelatase (hemH)
NMB0632	Iron(III) ABC transporter, ATP-binding protein (fbpB)	NMB0750	Bacterioferritin comigratory protein (bcp)
NMB0633	Iron(III) ABC transporter, permease protein	NMB1346	TonB-dependent receptor, putative, authentic frame shift
NMB0634	Iron(III) ABC transporter, periplasmic binding protein (fbpA)	NMB1449	TonB-dependent receptor, authentic point mutation
NMB1206	Bacterioferritin B (bfrB)	NMB1829	TonB-dependent receptor
NMB1207	Bacterioferritin A (bfrA)	NMB1882	TonB-dependent receptor
NMB1540	Lactoferrin-binding protein A (lbpA)	NMB1990	Iron(III) ABC transporter, permease protein
NMB1541	Lactoferrin-binding protein B (lbpB)	NMB1991	Iron(III) ABC transporter, permease protein
NMB1668	Hemoglobin receptor (hmbR)	NMB2132	Transferrin-binding protein-related protein
NMB1880	ABC transporter, periplasmic solute-binding protein, putative		
NMB1989	Iron(III) ABC transporter, periplasmic binding protein		
		Colonization	
NMB0018	Pilin PilE	NMB0014	3-Deoxy-D-manno-octulosonic-acid transferase (kdtA)
NMB0019	pilS cassette	NMB0181	Outer membrane protein OmpH, putative
NMB0020	pilS cassette	NMB0547	Type IV pilin protein
NMB0021	pilS cassette	NMB0586	Adhesin, putative
NMB0022	pilS cassette	NMB0992	Adhesin (hsf)
NMB0023	pilS cassette	NMB0995	Macrophage infectivity potentiator-related protein
NMB0024	pilS cassette	NMB1483	Lipoprotein NlpD, putative
NMB0025	Large pilS cassette	NMB1567	Macrophage infectivity potentiator
NMB0026	pilS cassette	NMB1946	Outer membrane lipoprotein
NMB0049	pilC2 protein, authentic frame shift	NMB1961	VacJ-related protein
NMB0051	Twitching motility protein	NMB1994	Adhesin/invasin, putative
NMB0052	Twitching motility protein PilT		
NMB0182	Outer membrane protein Omp85		
NMB0329	Type IV pilus assembly protein (pilF)		
NMB0332	Type IV prepilin peptidase (pilD)		
NMB0333	Pilus assembly protein PilG		
NMB0382	Outer membrane protein class 4 (rmpM)		
NMB0442	Opacity protein, authentic frame shift		
NMB0922	α-2,3-Sialyltransferase		
NMB0926	Opacity protein		
NMB1053	Class 5 outer membrane protein (opc)		
NMB1429	Outer membrane protein PorA		
NMB1465	Opacity protein, authentic frame shift		
NMB1527	ADP-heptose–LPS heptosyltransferase II (rfaF)		
NMB1636	Opacity protein, authentic frame shift		
NMB1704	β-1,4-glucosyltransferase (lgtF)		
NMB1705	α-1,2-*N*-acetylglucosamine transferase (rfaK)		
NMB1926	Lacto-*N*-neotetraose biosynthesis glycosyl transferase LgtE		
NMB1928	Lacto-*N*-neotetraose biosynthesis glycosyl transferase LgtB		
NMB1929	Lacto-*N*-neotetraose biosynthesis glycosyl transferase LgtA		
NMB1985	Adhesion and penetration protein (hap)		
NMB2032	Lipopolysaccharide glycosyl transferase, authentic frame shift (lgtG)		
NMB2039	Major outer membrane protein PIB (porB)		
NMB2156	Heptosyltransferase I (rfaC)		
		Evasion	
NMB0067	Polysialic acid capsule biosynthesis protein SiaD, truncation	NMB0393	Multidrug resistance protein
NMB0068	Polysialic acid capsule biosynthesis protein SiaC	NMB2127	Protease, putative
NMB0069	Polysialic acid capsule biosynthesis protein SiaB		
NMB0070	Polysialic acid capsule biosynthesis protein synX		
NMB0071	Capsule polysaccharide export outer membrane protein CtrA		

Table 1. Continued.

Previously known genes		New genes indentified in strain MC58	
NMB0072	Capsule polysaccharide export inner-membrane protein CtrB		
NMB0073	Capsule polysaccharide export inner-membrane protein CtrC		
NMB0074	Capsule polysaccharide export ATP-binding protein CtrD		
NMB0082	Capsule polysaccharide modification protein LipA		
NMB0083	Capsule polysaccharide modification protein LipB		
NMB0216	Catalase (kat)		
NMB0700	IgA-specific serine endopeptidase		
NMB1332	Carboxy-terminal peptidase (prc)		
NMB1398	Cu-Zn-superoxide dismutase (sodC)		
NMB1714	Multidrug efflux pump channel protein MtrE		
NMB1715	Multiple transferable resistance system protein MtrD		

Toxins

Previously known genes		New genes indentified in strain MC58	
NMB0017	UDP-3-O-hydroxymyristoyl N-acetylglucosamine deacetylase (envA)	NMB0178	Acyl-(acyl-carrier-protein)–UDP–N-acetylglucosamine O-acetyltransferase (lpxA)
NMB0199	Lipid-A-disaccharide synthase (lpxB)	NMB0179	(3R)-Hydroxymyristoyl-(acyl carrier protein) dehydratase (fabZ)
NMB0493	Hemagglutinin/hemolysin-related protein	NMB0180	UDP-3-O-(3- hydroxymyristoyl)-glucosamine N-acetyltransferase (lpxD)
NMB0496	Hemolysin activator-related protein	NMB1210	Toxin-activating protein, putative
NMB0585	Iron-regulated protein FrpA, putative	NMB1646	Hemolysin, putative
NMB1214	Hemagglutinin/hemolysin-related protein	NMB1762	Hemolysin activation protein HecB, putative
NMB1407	FrpA-related protein, degenerate	NMB1763	Toxin-activating protein, putative
NMB1415	Iron-regulated protein FrpC		
NMB1768	Hemagglutinin/hemolysin-related protein		
NMB1779	Hemagglutinin/hemolysin-related protein		
NMB1780	Hemolysin activation protein HecB, putative		

Unknown

		NMB1753	VapD-related protein

Table 2. Potentially phase-variable genes in *N. meningitidis* strain MC58, listed by functional category.

Functional category	Previously known	New
Outer-membrane proteins	6: NMB0442, NMB0926, NMB1053, NMB1429, NMB1465, NMB1636	6: NMB0182, NMB1277, NMB1985, NMB1994, NMB1998, NMB2104
LPS	2: NMB1929, NMB2032	2: NMB0624, NMB1836
Pilus	3: NMB0049, NMB0218, NMB1847	0
Iron acquisition	1: NMB1668	2: NMB1540, NMB1988
Toxins and related	0	2: NMB1407, NMB1969
Capsule	1: NMB0067	0
Restriction/modification systems	0	6: NMB0726, NMB0831, NMB1032, NMB1223, NMB1261, NMB1375
Bacteriophage/ bacteriocin	0	4: NMB0098, NMB0961, NMB1080, NMB1783
Others	0	5: NMB1140, NMB1379, NMB1723, NMB1734, NMB1913
Conserved hypothetical proteins	2: NMB0312, NMB1525	12: NMB0415, NMB0432, NMB0471, NMB0486, NMB0593, NMB1265, NMB1543, NMB0970, NMB1634, NMB1741, NMB1760, NMB1893
Hypothetical proteins	1: NMB1507	10: NMB0032, NMB0065, NMB0300, NMB0368, NMB0488, NMB1275, NMB1489, NMB1786, NMB1931, NMB2008
Total	16	49

Another mechanism that contributes to the evasion of the host immune system in meningococcus is phase variation that controls gene expression. Indeed, most of the recognized host-interactive factors are phase variable (*48*); these have been called "contingency loci" (*49*). The genetic basis for this variation depends on the evolution of iterative DNA motifs, especially homopolymeric tracts, to effect reversible, high-frequency molecular switching through slippage-like mechanisms (*49*). Although the mutation rate at these loci is high in all strains of *N. meningitidis*, pathogenic strains may be hypermutable at these loci due in part to genetic defects in the DNA mismatch repair process (*50*). The genome sequence shows that strain MC58 has a defect in the Dam methylase similar to those found in the hypermutable strains. The repeats typically associated with phase variation and their contexts have been analyzed (*51*) to reveal all potentially phase-variable genes in strain MC58 (Table 2). All genes previously recognized to be phase variable in meningococci were identified in strain MC58. The repertoire differs between strain MC58 and other reported strains, e.g., in the iron acquisition and LPS biosynthetic genes. A number of novel phase-variable surface proteins, restriction modification systems, and genes not previously associated with repeat mediated phase variation, e.g., toxin secretion systems, have been identified, as well as several hypothetical proteins. Genome analysis shows that strain MC58 has a far greater number of putative phase-variable genes than have been recognized in other organisms studied to

date. In particular, *H. influenzae* contains only 14 identified phase-variable genes, most of which are involved in virulence.

The complete genome sequence of *N. meningitidis* strain MC58 provides a new starting point for the study of the pathogenesis of meningitis. The identification of a large number of putative virulence determinants that were not previously identified in meningococci and the unprecedented number of phase-variable genes provide new insights into potential mechanisms of the pathogenesis of invasive disease. On the basis of the genome sequence of strain MC58, several long-awaited candidates for vaccination against the serogroup B meningococcus have been identified and characterized experimentally (*11*). The development of a vaccine against *H. influenzae* significantly reduced the incidence of infections caused by this organism. A vaccine against the serogroup B meningococcus holds similar promise for dramatically reducing the morbidity and mortality of *Neisseria*-mediated meningitis.

References and Notes

1. B. Schwartz, P. S. Moore, C. V. Broome, *Clin. Microbiol. Rev.* **2** (Suppl.), S118 (1989).
2. E. C. Gotschlich, T. Y. Liu, M. S. Artenstein, *J. Exp. Med.* **129**, 1349 (1969); E. C. Gotschlich, I. Goldschneider, M. S. Artenstein, *J. Exp. Med.* **129**, 1367 (1969).
3. D. Jones, in *Meningococcal Disease*, K. Cartwright, Ed. (Wiley, New York, 1995), p. 147.
4. B. T. McGuinness *et al.*, *Lancet* **337**, 514 (1991).
5. Sequence data for *N. meningitidis* strain Z2491 were obtained from The Sanger Centre Web site at www.sanger.ac.uk/Projects/N_meningitidis/. Sequencing of *N. meningitidis* strain Z2491 was accomplished with support by The Wellcome Trust.
6. R. D. Fleischmann *et al.*, *Science* **269**, 496 (1995).
7. Cloning, sequencing, and assembly were as described for previous genomes sequenced by TIGR [R. D. Fleischmann *et al.*, *Science* **269**, 496 (1995); C. M. Fraser *et al.*, *Science* **270**, 397 (1995); C. J. Bult *et al.*, *Science* **273**, 1058 (1996); C. M. Fraser *et al.*, *Nature* **390**, 580 (1997); J.-F. Tomb *et. al.*, *Nature* **388**, 539 (1997); H. P. Klenk *et al.*, *Nature* **390**, 364 (1997); C. M. Fraser *et al.*, *Science* **281**, 375 (1998); M. J. Gardner *et al.*, *Science* **282**, 1126 (1998); K. E. Nelson *et al.*, *Nature* **399**, 323 (1999); O. White *et al.*, *Science* **286**, 1571 (1999)]. One small-insert library (1.5 to 2.5 kb) was generated by random mechanical shearing (nebulization) of genomic DNA and cloning into a pUC-derived vector. One large-insert lambda library (8 to 24 kb) was generated by partial Sau 3A digestion of genomic DNA and ligation to λ-DASHII/ Bam HI vector (Stratagene). In the initial random-sequencing phase, 8.4-fold sequence coverage was achieved with 37,201 sequences from plasmid clones (average read length: 515 nucleotides) and 3.2-fold clone coverage was obtained with 880 sequences from λ clones as a scaffold. The plasmid and λ sequences were jointly assembled into 193 contigs larger than 1.5 kb with TIGR Assembler [G. S. Sutton, O. White, M. D. Adams, A. R. Kerlavage, *Genome Sci. Technol.* **1**, 9 (1995)]. The coverage criteria were that every position required at least double-clone coverage [or sequence from a polymerase chain reaction (PCR) product amplified from genomic DNA], and either sequence from both strands or with two different sequencing chemistries. The sequence was edited manually with the TIGR Editor, and additional PCR and sequencing reactions were performed to close gaps, improve coverage, and resolve sequence ambiguities. The final genome sequence is based on 40,020 sequences.
8. S. L. Salzberg, A. J. Salzberg, A. R. Kerlavage, J. F. Tomb, *Gene* **217**, 57 (1998).
9. J. R. Lobry, *Mol. Biol. Evol.* **13**, 660 (1996).
10. The first annotation method was based on the prediction of ORFs likely to encode proteins by means of Glimmer [S. L. Salzberg, A. L. Delcher, S. Kasif, O. White, *Nucleic Acids Res.* **26**, 544 (1998)]. This program, based on interpolated Markov models, was trained with ORFs larger than 600 bp from the genomic sequence, as well as with the *N. meningitidis* genes available in GenBank. All predicted ORFs were searched against a nonredundant amino acid database as described [R. D. Fleischmann *et al.*, *Science* **269**, 496 (1995)]. Protein membrane-spanning domains were identified by TopPred [M. G. Claros and G. von Heijne, *Comput. Appl. Biosci.* **10**, 685 (1994)]. In the second annotation approach, the whole-genome sequence was directly used for homology searches with BLASTX, BLASTN, and TBLASTX [S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990); S. F. Altschul *et al.*, *Nucleic Acids Res.* **25**, 3389 (1997)], without introducing assumptions associated with defined ORFs. Both a Java-based graphical interface (TIGR software, unpublished) and an ACEDB graphical interface [R. Durbin and J. T. Mieg, a *C. elegans* database (1991), documentation, code, and data available from anonymous FTP servers at lirmm.lirmm.fr, cele.mrc-lmb. cam.ac.uk and ncbi.nlm.nih.gov; N. J. Saunders, J. F. Peden, D. W. Hood, E. R. Moxon, *Mol. Microbiol.* **27**, 1091 (1998)] were used to visualize all potential ORFs, homologies, and repeats and to perform local sequence interrogation. The 5′ regions of each ORF were inspected to define initiation codons by using homologies, position of ribosomal binding sites, putative promoter sequences, and transcriptional terminators.
11. M. Pizza *et al.*, *Science* **287**, 1816 (2000).
12. Supplemental Web material is available at *Science* Online at www.sciencemag.org/feature/data/1046515.shl.
13. M. Riley, *Microbiol. Rev.* **57**, 862 (1993).
14. Two sets of hidden Markov models (HMMs) were used to determine ORF membership in families and superfamilies. These included 1407 HMMs from pfam v4.0 and 500 HMMs from the TIGR functionally conserved protein families.
15. Likely recent duplication events were inferred by identifying ORFs that were most similar to another ORF from the *N. meningitidis* genome relative to ORFs from all other complete genome sequences. All ORFs were searched with FASTA3 against all ORFs from the complete genomes of yeast, six Archaea, and 20 bacteria, as well as all available ORFs from *C. elegans*. Matches with a FASTA *p* value of 10^{-5} were considered significant.
16. J. Mahillon and M. Chandler, *Microbiol. Mol. Biol. Rev.* **62**, 725 (1998).
17. J. S. Kroll, B. M. Loynds, E. R. Moxon, *Mol. Microbiol.* **5**, 1549 (1991); S. R. Dobson, J. S. Kroll, E. R. Moxon, *Infect. Immun.* **60**, 618 (1992).
18. A transposase identified in the whole-genome sequence of *Syncehocystis* sp. PCC6803. GenBank accession number D90911 GI:1653129.
19. J. L. Rasmussen, D. A. Odelson, F. L. Macrina, *J. Bacteriol.* **169**, 3573 (1987).
20. A. I. Knight, H. Ni, K. A. Cartwright, J. J. McFadden, *Mol. Microbiol.* **6**, 1565 (1992); P. Zhu, G. Morelli, M. Achtman, *Mol. Microbiol.* **33**, 635 (1999).
21. C. S. Carrick, J. A. Fyfe, J. K. Davies, *Gene* **220**, 21 (1998).
22. C. F. Marrs, F. W. Rozsa, M. Hackel, S. P. Stevens, A. C. Glasgow, *J. Bacteriol.* **172**, 4370 (1990).
23. S. D. Goodman and J. J. Scocca, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 6982 (1988).
24. All ORFs were searched with FASTA3 against all ORFs from the complete genomes of yeast, six Archaea, and 20 bacteria, as well as all available ORFs from *C. elegans*. Matches with a FASTA *p* value of 10^{-5} were considered significant.
25. R. A. Alm *et al.*, *Nature* **397**, 176 (1999).
26. R. P. Anderson and J. R. Roth, *Cold Spring Harbor Symp. Quant. Biol.* **43**, 1083 (1978); C. W. Hill and J. A. Gray, *Genetics* **119**, 771 (1988); S. L. Liu and K. E. Sanderson, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 10303 (1996).
27. J. G. Lawrence and H. Ochman, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 9413 (1998).
28. ———, *J. Mol. Evol.* **44**, 383 (1997).
29. Dinucleotide signatures analysis: The data were computed as described [S. Karlin, A. M. Campbell, J. Mrazek, *Annu. Rev. Genet.* **32**, 185 (1998)], with a window size of 100,000 bp and a granularity of 100 bp. χ^2 analysis: The distribution of all 64 trinucleotides (3mers) was computed for the complete genome in all six reading frames, followed by the 3mer distribution in 20,000-bp windows. Windows overlapped by 10,000 bp. For each window, the χ^2 statistic on the difference between its 3mer content and that of the whole genome was computed.
30. The best database match was to an unpublished gene sequence described as "Identification and analysis of TspB, a potent T-cell and B-cell stimulating *Neisseria*-specific antigen." Submitted by D. A. A. Ala'Aldeen. GenBank accession number AJ010115 GI:3758869.
31. The best database match was to an unpublished gene sequence described as "Structural analysis of putative hypovirulent plasmid, pJTPS1, found in a spontaneous avirulent mutant of *Rastonia solanacearum*." Submitted by R. Shimizu *et al.* GenBank accession number AB015669 GI:3402502.
32. P. Cao and T. L. Cover, *J. Bacteriol* **179**, 2852 (1997).
33. F. F. Correia, S. Inouye, M. Inouye, *J. Bacteriol.* **167**, 1009 (1986); *J. Biol. Chem* **263**, 12194 (1988).
34. J. Hacker, G. Blum-Oehler, I. Muhldorfer, H. Tschape, *Mol. Microbiol.* **23**, 1089 (1997).
35. H. Neubauer, E. Glaasker, W. P. Hammes, B. Poolman, W. N. Konings, *J. Bacteriol.* **176**, 3007 (1994); M. A. Ehrmann and R. F. Vogel, *FEMS Microbiol. Lett.* **169**, 81 (1998).
36. E. Holten, *Acta Pathol. Microbiol. Scand. B* **84**, 9 (1976).
37. S. A. Morse and B. H. Hebeler, *Infect. Immun.* **21**, 87 (1978).
38. C. N. Cornelissen and P. F. Sparling, *Mol. Microbiol.* **14**, 843 (1994); A. B. Schryvers and I. Stojiljkovic, *Mol. Microbiol.* **32**, 1117 (1999).
39. K. G. Wooldridge and P. H. Williams, *FEMS Microbiol. Rev.* **12**, 325 (1993).
40. H. Yoon and B. M. Anderson, *J. Gen. Microbiol.* **135**, 245 (1989).
41. M. Virji *et al.*, *Mol. Microbiol.* **5**, 1831 (1991); C. F. Rayner, A. Dewar, E. R. Moxon, M. Virji, R. Wilson, *J. Infect. Dis.* **171**, 113 (1995).
42. X. Nassif *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 3769 (1994).
43. R. Haas and T. F. Meyer, *Cell* **44**, 107 (1986); R. Haas, S. Veit, T. F. Meyer, *Mol. Microbiol.* **6**, 197 (1992).
44. L. A. Wainwright, K. H. Pritchard, H. S. Seifert, *Mol. Microbiol.* **13**, 75 (1994); L. A. Wainwright, J. V. Frangipane, H. S. Seifert, *Nucleic Acids Res.* **25**, 1362 (1997).
45. J. L. Diaz, M. Virji, J. E. Heckels, *FEMS Microbiol. Lett.* **21**, 181 (1984); R. W. Pinner, P. A. Spellman, D. S. Stephens, *Infect. Immun.* **59**, 3169 (1991).
46. H. J. Jennings, A. Gamian, F. E. Ashton, *J. Exp. Med.* **165**, 1207 (1987); M. P. Jennings, D. W. Hood, I. R. Peak, M. Virji, E. R. Moxon, *Mol. Microbiol.* **18**, 729 (1995); N. J. Klein *et al.*, *J. Infect. Dis.* **173**, 172 (1996).
47. M. Frosch, C. Weisgerber, T. F. Meyer, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 1669 (1989); M. Frosch, U. Edwards, K. Bousset, B. Krausse, C. Weisgerber, *Mol. Microbiol.* **5**, 1251 (1991).
48. N. J. Saunders, A. C. Jeffries, J. F. Peden, D. W. Hood, E. R. Moxon, in preparation.
49. E. R. Moxon, P. B. Rainey, M. A. Nowak, R. E. Lenski, *Curr. Biol.* **4**, 24 (1994).
50. C. Bucci *et al.*, *Mol. Cell* **3**, 435 (1999).
51. N. J. Saunders, J. F. Peden, D. W. Hood, E. R. Moxon, *Mol. Microbiol.* **27**, 1091 (1998).
52. We thank M. Heaney, J. Scott, M. Holmes, D. Richardson, B. Loftus, B. Lee, and B. Vincent for software and database support; the TIGR faculty and sequencing core for advice and assistance; the Oxford Super Computing Centre for access to its resources; and D. Kingsbury for initial handling of sequences for Chiron. Supported by Chiron Corporation.

25 October 1999; accepted 27 January 2000