Journal of Bacteriology, July 2011, p. 3633–3641 Vol. 193, No. 14
0021-9193/11/$12.00 doi:10.1128/JB.01198-10

Using Single-Nucleotide Polymorphisms To Discriminate Disease-Associated from Carried Genomes of *Neisseria meningitidis*▽†

Lee S. Katz,[1,2]‡ Nitya V. Sharma,[1]‡ Brian H. Harcourt,[2] Jennifer Dolan Thomas,[2] Xin Wang,[2] Leonard W. Mayer,[2] and I. King Jordan[1,3]*

School of Biology, Georgia Institute of Technology, Atlanta, Georgia 30332[1]; Meningitis and Vaccine Preventable Diseases Branch, Centers for Disease Control and Prevention, Atlanta, Georgia 30333[2]; and PanAmerican Bioinformatics Institute, Santa Marta, Magdalena, Colombia[3]

Received 6 October 2010/Accepted 2 May 2011

Neisseria meningitidis **is one of the main agents of bacterial meningitis, causing substantial morbidity and mortality worldwide. However, most of the time *N. meningitidis* is carried as a commensal not associated with invasive disease. The genomic basis of the difference between disease-associated and carried isolates of *N. meningitidis* may provide critical insight into mechanisms of virulence, yet it has remained elusive. Here, we have taken a comparative genomics approach to interrogate the difference between disease-associated and carried isolates of *N. meningitidis* at the level of individual nucleotide variations (i.e., single nucleotide polymorphisms [SNPs]). We aligned complete genome sequences of 8 disease-associated and 4 carried isolates of *N. meningitidis* to search for SNPs that show mutually exclusive patterns of variation between the two groups. We found 63 SNPs that distinguish the 8 disease-associated genomes from the 4 carried genomes of *N. meningitidis*, which is far more than can be expected by chance alone given the level of nucleotide variation among the genomes. The putative list of SNPs that discriminate between disease-associated and carriage genomes may be expected to change with increased sampling or changes in the identities of the isolates being compared. Nevertheless, we show that these discriminating SNPs are more likely to reflect phenotypic differences than shared evolutionary history. Discriminating SNPs were mapped to genes, and the functions of the genes were evaluated for possible connections to virulence mechanisms. A number of overrepresented functional categories related to virulence were uncovered among SNP-associated genes, including genes related to the category "symbiosis, encompassing mutualism through parasitism."**

Neisseria meningitidis, a meningococcus, is a leading cause of bacterial meningitis worldwide, with devastating morbidity and mortality (Centers for Disease Control and Prevention [http://www.cdc.gov/meningitis/about/faq.html]). *N. meningitidis* is a Gram-negative encapsulated diplococcus of the human nasopharynx that is most frequently found to be asymptomatically carried (27); ~10% of healthy individuals are carriers of *N. meningitidis* (9, 12). In a small number of carriers, some strains of *N. meningitidis* are able to invade epithelium of the nasopharynx and enter the bloodstream, thus leading to invasive disease, such as meningococcal meningitis and meningococcemia (38).

A number of previous studies have taken a comparative genomics approach to try to determine if there is any genomic basis for the difference between disease-associated and asymptomatically carried isolates of *N. meningitidis*. Perrin et al. used comparative genome hybridization (CGH) to compare the genomes of disease-associated *N. meningitidis* strains against the genomes of the closely related species *Neisseria gonorrhoeae* and *Neisseria lactamica* (33). They were able to find a number of chromosomal regions present only in *N. meningitidis*, suggesting a possible role in species-specific virulence. However, genes found in the species-specific regions would later be shown to be present in both disease-associated and carried isolates of *N. meningitidis* (42). A subsequent CGH study discovered 55 genes present in all *N. meningitidis* serogroup B isolates analyzed and absent in *Neisseria* commensal species (46). Nevertheless, several of these serogroup B strains were carried isolates that were not associated with disease. It was also shown later that the majority of genes previously implicated in virulence using the comparative approach were shared between *N. meningitidis* and the nonpathogenic *N. lactamica* (45).

In 2005, Bille et al. discovered an 8-kb bacteriophage-derived sequence shared among 29 disease-associated *N. meningitidis* genomes and largely absent from carried isolates (5). While no single gene in this prophage met the condition of being present in all disease-associated genomes and absent in all carried genomes, the distribution of prophage genes was highly skewed toward disease-associated genomes. Thus, at that time, this genetic island represented the best example of a genomic feature that could distinguish disease-associated genomes from carried genomes of *N. meningitidis*. However, the next year, Hotopp et al. used a more exhaustive CGH study to show that this prophage was actually found in the genomes of 60% of disease-associated isolates and 42% of carried isolates (16).

* Corresponding author. Mailing address: School of Biology, Georgia Institute of Technology, Atlanta, GA 30332. Phone: (404) 385-2224. Fax: (404) 894-0519. E-mail: king.jordan@biology.gatech.edu.

† Supplemental material for this article may be found at http://jb.asm.org/.

‡ These authors contributed equally to the work.

▽ Published ahead of print on 27 May 2011.

In 2008, Schoen et al. performed the first complete genome sequence-based comparison between disease-associated and carried isolates of *N. meningitidis* (42). These authors found that genes previously implicated in virulence were widely shared among disease-associated and carried genomes; in other words, there does not appear to be any core pathogenome for *N. meningitidis*. Later, the same group used genome sequence comparison between a disease-associated and a carried isolate of serogroup B *N. meningitidis* strains to show that virulence in *N. meningitidis* is likely to be encoded by sequence differences found across numerous genes (20). Taken together, the results of all of these comparative genomic studies indicate that the presence or absence of specific genes, sets of genes, or other large-scale genomic features cannot be used to distinguish disease-associated isolates from carried isolates of *N. meningitidis*.

In light of these previous results, we decided to explore the utility of individual nucleotide variations for discriminating between disease-associated and carried isolates of *N. meningitidis*. The use of nucleotide variation takes advantage of advances in sequencing technology to provide a deeper level of resolution for genome comparisons. We hypothesized that single nucleotide polymorphisms (SNPs) will provide markers that can distinguish disease-associated isolates from carried isolates of *N. meningitidis*. To test this hypothesis, we compared complete genome sequences of 8 disease-associated and 4 carried isolates of *N. meningitidis* (4, 23, 31, 32, 42, 48). The designation of isolate genomes as disease associated or carried was based on two factors. First, the disease-associated isolates were sampled from individuals with meningococcal disease, and the carried isolates were taken from asymptomatic individuals. Second, carried isolates represent sequence types (STs) that are only very infrequently or never associated with disease. In this way, the two groups of isolates represent instances of phenotypic differences in *N. meningitidis* virulence, and we sought to assess whether there may be genomic determinants of these differences. To do this, we searched for SNPs that show mutually exclusive patterns of variation between the two groups of isolates. We found that tens of SNPs can serve as markers that distinguish these sets of disease-associated and carried isolates of *N. meningitidis*, and these discriminating SNPs are more likely to reflect phenotypic differences than shared evolutionary history. We mapped these discriminating SNPs to *N. meningitidis* genes to assess their potential functional significance.

The transition from asymptomatic to disease-associated states for *N. meningitidis* may be extremely rapid. Thus, the accumulation of discriminating SNPs observed here probably does not represent real-time genetic changes but rather represents combinations of existing SNPs from standing genetic variation, perhaps introduced via recombination, that may either serve to identify lineages with invasive potential or predispose strains to the invasive state. In addition, some of the carried isolates studied here belong to serogroup and ST combinations that have been shown to cause a small percentage of meningococcal disease, and isolates from disease-associated strains spend most of their time being carried. In other words, the disease-associated and carried isolates studied here yield a snapshot in time and place of a set of nucleotide variants that distinguish one group of *N. meningitidis* disease-associated isolates from a group of carried isolates. Accordingly, the particular set of discriminating SNPs characterized here, along with the list of SNP-associated genes, may change as additional genome sequences are characterized and compared.

MATERIALS AND METHODS

***N. meningitidis* culture conditions and DNA extraction.** Isolates were stored at −80°C in defibrinated sheep blood (Lampire, Pipersville, PA) prior to use and were subsequently streaked onto chocolate II agar (BBL, Sparks, MA) and incubated at 37°C overnight with 5% CO_2 before being harvested for DNA preparation. Purified genomic DNA was extracted using the blood and cell culture DNA maxikit (Qiagen, Valencia, CA) by following the manufacturer's instructions. The DNA concentration and the 260/280 ratio were obtained using a NanoDrop ND-1000 spectrophotometer (NanoDrop Products, Wilmington, DE).

Genome sequencing and analysis. Sequencing of *N. meningitidis* isolates M9261, M13220, M10699, M17062, and M15141 was performed using Roche Applied Science/454 pyrosequencing in the CDC Biotechnology Core Facility; each strain was sequenced using the GS-20 platform, with the exception of M17062, which was sequenced using the GS-FLX Titanium platform. For each genomic DNA preparation, a random shotgun library was produced using Roche protocols for nebulization, end polishing, adaptor ligation, nick repair, and single-stranded library formation (28). After emulsion PCR, DNA-bound beads were isolated and sequenced using long read (LR) sequencing kits. Sequencing was followed by read trimming and refiltering to recover short quality reads.

Genome sequence assemblies were performed using a customized genome analysis pipeline (CG-Pipeline version 0.2.1) (23) that combines reference-based assembly using the Newbler assembler and AMOScmp (34). Results from the two assemblers were combined using Minimus. The CG-Pipeline platform was also used to perform gene prediction and functional annotation using a combination of tools. Additionally, PromPredict and TransTermHP were used to predict promoter and terminator regulatory regions, respectively (22, 37). The four annotated genomes (see Table 2) were uploaded into a customized database, *Neisseria* Base (NBase), based on the GBrowse platform (47).

Genome alignment and SNP analysis. The five *N. meningitidis* isolates characterized here were analyzed together with 7 other completely sequenced *N. meningitidis* isolates (Table 1). Complete genome sequences were aligned using the program MAUVE, which locates and aligns conserved and syntenic genomic regions called local colinear blocks (LCBs) (10). Default MAUVE alignment settings were used except for a minimum LCB weight of 500. M13220 was set as the reference genome, whereby all other genomes were rearranged according to M13220. Only LCBs that contained conserved regions of all 12 isolates ($n = 250$) were used for subsequent SNP analysis. The 250 individual LCBs were realigned using ClustalW (version 2.0.12) (49). LCB alignments were analyzed to look for discriminating nucleotide patterns (i.e., SNPs) separating the 8 disease-associated genomes from the 4 carried genomes (Table 1). The disease-associated isolates were sampled from individuals with meningococcal disease, and the carried isolates were taken from asymptomatic individuals. Here, SNPs include nucleotide variations along with insertions and deletions (indels). A discriminating SNP is defined as a polymorphic site that shows one nucleotide pattern for one group of genomes (disease associated or carried) and a mutually exclusive pattern for the other group (Fig. 1). The total numbers of disease-associated and carried SNPs were computed and compared to a null distribution of expected discriminating SNP counts, given the background polymorphism level, generated using simulation. For simulation, the identities of the nucleotides at each polymorphic site were randomly permuted among genomes. Polymorphic sites were then evaluated to come up with a count of discriminating SNPs for simulation. This process was repeated 10,000 times. Two additional controls consisting of discriminating SNPs followed by simulation analyses were done by comparing (i) only genomes within the disease-associated group and (ii) only genomes within the carried group. For the 8 disease-associated genomes, two groups were created: (i) all three serogroup C isolates plus the closely related ST-11 serogroup W135 isolate and (ii) all remaining invasive isolates from serogroups A and B. This was done in an attempt to create two evolutionarily distinct groups of invasive strains that may be expected to be most divergent with respect to their SNP profiles. Discriminating SNPs were identified between the groups, and a distribution of the expected number of discriminating SNPs was generated with random permutation of polymorphic sites as described above. Similarly, for the 4 asymptomatically carried isolate genomes, the two comparison groups consisted of (i) only nongroupable isolates (α14 and M17062) and (ii) only groupable isolates (α153 and α275).

TABLE 1. *N. meningitidis* genomes compared in this study[a]

Name	Serogroup	ST	CC	Isolate phenotype	Location (yr)	NCBI accession no.
M13220	A	ST-7	ST-5	Disease	Philippines (2005)	SRS074220
M10699	B	ST-32	ST-32	Disease	Oregon (2003)	SRS074512
M15141	C	ST-11	ST-11	Disease	New York (2006)	SRS074514
M9261	W135	ST-11	ST-11	Disease	Burkina Faso (2002)	SRS074515
053442	C	ST-4821	ST-4821	Disease	China (2003)	CP000381
Z2491	A	ST-4	ST-4	Disease	Gambia (1983)	AL157959
FAM18	C	ST-11	ST-11	Disease	North Carolina (1980s)	AM421808
MC58	B	ST-74	ST-32	Disease	United Kingdom (1985)	AE002098
α14	*cnl*	ST-53	ST-53	Carried	Bavaria (1999-2000)	AM889136
α153	29E	ST-60	ST-60	Carried	Bavaria (1999-2000)	AM889137
α275	W135	ST-22	ST-22	Carried	Bavaria (1999-2000)	AM889138
M17062	NG	ST-198	ST-198	Carried	Minnesota (2008)	SRS188701.1

[a] ST, sequence type; CC, clonal complex; *cnl*, capsule-null locus (8); NG, nongroupable.

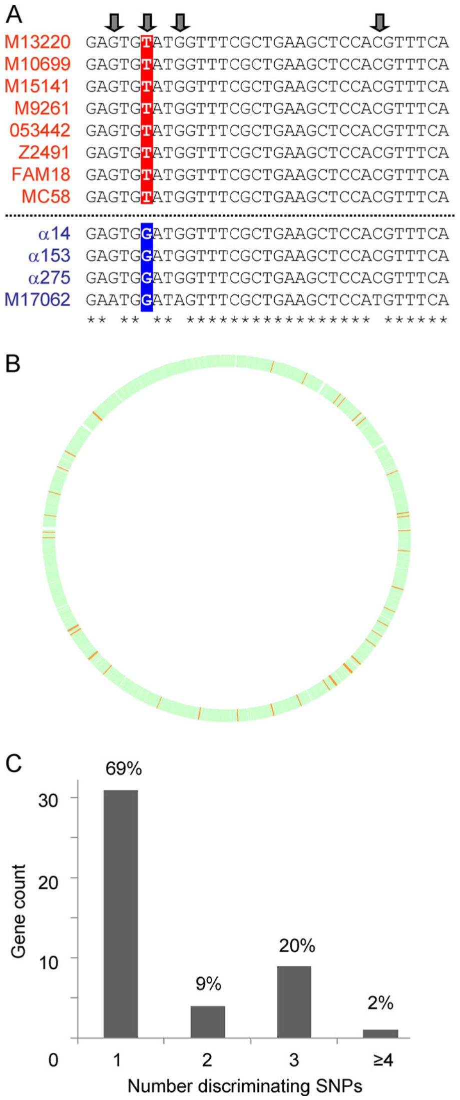

FIG. 1. SNPs that discriminate between disease-associated and carried isolate genomes. (A) Genome sequence alignment between disease-associated (red) and carried (blue) isolates was analyzed for the presence of SNPs (gray arrows). An example of a discriminating SNP for a disease-associated genome versus a carried genome is highlighted. (B) A concatenated genome map of M13220; genes are shown in green, and discriminating SNP locations are shown in orange. (C) A histogram of the number of discriminating SNPs per SNP-associated gene.

Phylogenetic analysis. Whole-genome sequence alignments were used to calculate nucleotide *p*-distances between *N. meningitidis* isolate genomes, and the distances were used to reconstruct an *N. meningitidis* phylogeny with the neighbor-joining algorithm (40) implemented in the program MEGA 4 (24). The same approach was used to reconstruct an *N. meningitidis* phylogeny based on a concatenated nucleotide sequence alignment of the 7 multilocus sequence typing (MLST) loci: *abcZ*, *adk*, *aroE*, *fumC*, *gdh*, *pdhC*, and *pgm* (15, 26). One thousand bootstrap replicates of the alignments were used to evaluate the confidence of the phylogenies.

F_{ST} analysis. The fixation index (F_{ST}) statistic (17) was used as a measure of the genetic differentiation between groups of *N. meningitidis* genomes. F_{ST} was measured using the pairwise nucleotide *p*-distances calculated from the *N. meningitidis* whole-genome alignment. Disease-associated and carried isolate groups were used to compute the average within-group genome *p*-distance (Ï*within*) and the average between-group genome *p*-distance (Ï*between*). F_{ST} was then calculated as 1 − (Ï*within*/Ï*between*). Simulation was used to compute a background distribution of F_{ST} values that could be expected given the levels of nucleotide variation among all genomes. To do this, *N. meningitidis* genomes were randomly assigned to either the disease-associated ($n = 8$) or carried ($n = 4$) groups, and F_{ST} was recalculated based on the random groups. This was repeated 10,000 times to yield a null frequency distribution of expected F_{ST} values.

Bayesian clustering method. We used a Bayesian method implemented by the program STRUCTURE version 2.3.1 to determine the optimal number of groups (*K*) that best represents the underlying nucleotide variation (i.e., the structure) among the *N. meningitidis* genomes analyzed here (35). STRUCTURE was run with $K = 1, 2, 3, 4$, and 5 groups. For each *K* value, the burn-in and run length parameter values were set to 50,000 each.

Genomic and functional characteristics of discriminating SNPs. Discriminating SNPs were mapped to *N. meningitidis* gene locations, either internal to or within 300 bp of the coding sequence. The resulting set of SNP-associated genes (proteins) was then evaluated for statistically significant enrichment for a variety of functional characteristics, including gene ontology (GO) annotations, the presence of a signal peptide, identity as a lipoprotein, horizontal transfer, subcellular location, and identity as a putative virulence factor. GO annotations were taken from the InterProScan database (52). The presence of signal peptides was inferred using the SignalP program (14). Lipoprotein status was inferred using the LipoP program (21). The horizontal transfer status of genes was inferred using a combination of three programs: BLAST (1), CodonO (2), and Alien-Hunter (50), along with an analysis of GC content. Genes were called as putative virulence factors based on the Virulence Factors Database (VFDB) (7, 51). SNP-associated genes were evaluated for statistical overrepresentation for each functional characteristic by using the hypergeometric test implement in the GeneMerge program (6). The hypergeometric test in this study gives the probability *P* of selecting *r* genes with a functional characteristic in the set of SNP-associated genes *k* from an overall set of genes in the genome *n*, where *p* is the proportion of *r* genes in the population and sampling is without replacement (equation 1). SNP-associated genes found to encode significantly overrepresented functions were further evaluated using BLAST homology searches from three sources: our genome browser NBase (http://nbase.biology.gatech.edu), NCBI's RefSeq database (36), and the NeMeSys database (39).

$$P(r|n,p,k) = \frac{C_r^{pn} C_{k-r}^{(1-p)n}}{C_k^n} \quad (1)$$

TABLE 2. New *N. meningitidis* genomes recently characterized[a]

Name	Location and date	No. of contigs	Assembly length (Mb)	No. of genes
M13220	Philippines, January 2005	82	2.20	2,108
M10699	Oregon, May 2003	40	2.18	1,978
M15141	New York City, August 2006	50	2.28	2,141
M9261	Burkina Faso, April 2002	79	2.21	2,261
M17062[b]	Minnesota	565	2.53	3,565

[a] These genomes, except M17062, are reported in their final form in reference 23.

[b] The assembly of M17062 is fragmented, likely due to the 454 multiplexing process. Due to the fragmentation, its true genome size is likely to be smaller than the reported assembly size, and the true number of genes is likely to be fewer.

RESULTS AND DISCUSSION

Comparative genomic sequence analysis of *N. meningitidis*. We hypothesize that SNPs can be used as markers that distinguish sets of disease-associated genomes from carried genomes of *N. meningitidis*. To test this hypothesis, and to search for potential genomic influences on virulence, we performed comparative sequence analysis of 12 isolates of *N. meningitidis* with completely sequenced genomes: 8 disease-associated isolates from individuals with meningococcal disease, and 4 carried isolates from asymptomatic individuals (Table 1). The 4 previously published disease-associated genomes are taken from a series of individual genome projects and represent the most common disease-associated *N. meningitidis* serogroups: A, B, and C (4, 31, 32, 48). The 3 previously published carried genomes were reported in 2008 as part of a comparative analysis of disease-associated and asymptomatically carried *N. meningitidis* genomes that focused on differences in the presence and absence of virulence factor genes between the two groups (42). Although these three isolates were taken from asymptomatic individuals during a carriage study (9), other isolates with the same serogroup and ST combination as α153 and α275 have been shown to cause a small percentage of meningococcal disease in Bavaria (42). Recently, we reported the characterization of 4 additional disease-associated genomes of *N. meningitidis* that cover serogroups A, B, and C and also include the first reported disease-associated W135 serogroup genome sequence (23). The W135 isolate characterized here was isolated in Burkina Faso and was the cause of a major outbreak of bacterial meningitis at the 2000 Hajj (13, 25, 29). Here, we also report the first ST-198 genome isolated from an asymptomatic individual; ST-198 is estimated to account for almost no disease cases (0.2%) in the United States (Active Bacterial Core Surveillance, unpublished data).

The *N. meningitidis* genomes were characterized via pyrosequencing on the Roche 454 instrument (Table 2). The number of reads produced in the 5 experiments ranged from 197,000 to 605,000, and the average read lengths were 105 to 245 bp. All together, these data yielded 47.6 to 94.3 million bases per genome, amounting to 20 to 40× coverage for the ∼2.2-Mb *N. meningitidis* genomes. We developed customized genome assembly, gene prediction, and functional annotation pipelines to analyze these data (23). Our genome assembly procedure resulted in an order-of-magnitude decrease in the number of contigs produced by the Newbler assembler that ships with the 454 platform. Additionally, our feature prediction procedure predicted genes with an estimated sensitivity of >95% (23). All 5 of the new genomes reported here, along with custom annotations and tools for searching and comparative sequence analysis, are available at our genome browser database (NBase [http://nbase.biology.gatech.edu]).

Single nucleotide polymorphisms discriminate between disease-associated and carried genomes of *N. meningitidis*. Previous comparative genomic sequence analyses of disease-associated isolates versus carried isolates of *N. meningitidis* failed to turn up evidence of obvious genomic differences between the two groups based on the presence or absence of any particular genes. In order to evaluate the genomic basis of the difference between disease-associated and carried genomes here, we focused our analysis on differences at the level of individual nucleotide variation. To do this, we compared genomic sequences of 8 disease-associated and 4 carried isolates of *N. meningitidis* (Table 1). Whole genome sequences of the *N. meningitidis* isolates were aligned as described in the Materials and Methods. There are 250 long orthologous regions (local colinear blocks) conserved among all 12 genomes, and the total length of the *N. meningitidis* genome sequence alignment is 1,544,040 positions, with the vast majority of positions (1,437,475; 93%) being absolutely conserved.

We identified aligned positions that show variation among genomes of *N. meningitidis* as SNPs. These SNPs include positions with insertion/deletion (i.e., alignment gap) variation among genomes. There are a total of 106,565 SNPs in the whole-genome *N. meningitidis* sequence alignment. We characterized SNPs that discriminate between the genomes of disease-associated and carried isolates of *N. meningitidi*s as those with mutually exclusive nucleotide patterns, including gap characters, between the two sets of sequences (Fig. 1A). There are 63 such discriminating SNPs, and they are distributed across the entire *N. meningitidis* genome (Fig. 1B). Discriminating SNPs were associated with individual *N. meningitidis* genes if they were found in the coding region or within 300 bp of a gene. The frequency distribution of discriminating SNPs per gene shows that the majority of SNP-associated genes are characterized by only one, or very few, discriminating SNPs (Fig. 1C). Taken together, the genomic and frequency distributions of discriminating SNPs indicate that there is no systematic bias in how these SNPs are sampled in genomic and/or alignment space.

The SNPs that discriminate between disease-associated and carried isolates represent a catalog of individual nucleotide variation with potential implications for understanding the genomic basis of virulence in *N. meningitidis*. However, this study covers a limited set of genomes, and it is likely that when additional or different sets of genome sequences are compared, distinct sets of discriminating SNPs will be observed. Furthermore, with respect to the sequence data analyzed here, it is possible that we observe these 63 discriminating SNPs simply by chance alone given the large number of SNP positions found in the whole-genome alignment analyzed here (106,565). We performed a simulation analysis to evaluate the probability of observing 63 discriminating SNP positions by chance alone given the background sequence variation among

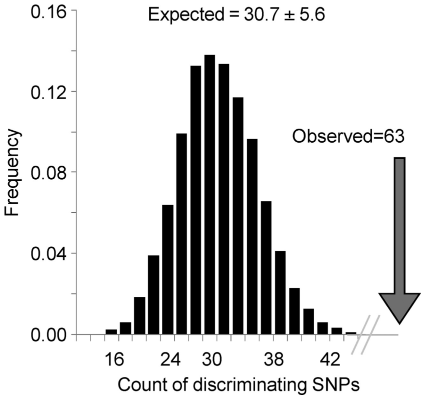

FIG. 2. Expected versus observed numbers of discriminating SNPs. The *N. meningitidis* genome sequence alignment was simulated to yield a null distribution of the expected number of discriminating SNPs given the background variation. The expected distribution is compared to the observed number of discriminating SNPs. See also Fig. S1 in the supplemental material.

the aligned genomes. To do this, the isolate identities of the aligned genomes were randomly permuted 10,000 times, and for each permutation, a number of discriminating SNPs over the permuted alignment was computed. This procedure resulted in a null distribution of discriminating SNP counts, parameterized against the actual background variation, against which we compared our observed value (Fig. 2). The value of the observed number of discriminating SNPs falls far outside the range of the entire set of simulated values. Accordingly, the observed number of discriminating SNPs is significantly greater than can be expected by chance alone given the background variation among the genomes of *N. meningitidis* studied here ($z = 5.73$, $P < 4.9e{-9}$, z test; or $P < 10e{-4}$ based on the simulation).

As an additional control, we performed two similar paired analyses of discriminating SNP detection and simulation by comparing only genomes within the disease-associated group and only genomes within the carried group. In each case, the control was set up to maximize genetic similarity within groups and genetic dissimilarity between groups as described in the Materials and Methods. In contrast to what was observed for the comparison of disease-associated and asymptomatically carried groups of genomes, we did not observe any overrepresentation of discriminating SNPs when genomes within the disease-associated group or genomes within the carried group were compared (see Fig. S1 in the supplemental material).

Discriminating polymorphisms reflect phenotype rather than shared evolutionary history. There are many more SNPs that discriminate between the disease-associated and carried genomes of *N. meningitidis* analyzed here than can be expected by chance alone. While these SNP data are suggestive of genomic differences with phenotypic relevance for virulence, they may also be attributed to shared evolutionary history. In other words, the abundance of discriminating SNPs may simply reflect the fact that the disease-associated isolates analyzed here are more closely related to each other than to the carried isolates and vice versa. If this is indeed the case, then the overall nucleotide sequence variation observed here should partition the disease-associated and carried isolates into two discrete groups of related genomes. We evaluated how *N. meningitidis* genome sequence variation is partitioned among the disease-associated and carried isolates studied here in several different ways: (i) using phylogenetic analyses to infer the evolutionary history of the genomes, (ii) using a standard population genetic measure—the fixation index (F_{ST})—to evaluate how nucleotide variation is partitioned within and between the disease-associated and carried groups, and (iii) using naive Bayesian clustering of the observed SNP variation.

We reconstructed the phylogenies of the *N. meningitidis* genomes analyzed here in order to assess their evolutionary relationships. Specifically, we sought to evaluate whether the disease-associated and carried genomes form distinct phylogenetic groups, each of which shares a unique common ancestor (i.e., monophyletic clades). To do this, we first aligned the seven housekeeping loci used for multilocus sequence typing (MLST) (15, 26) among the 11 genomes analyzed here and reconstructed a phylogeny based on the concatenated alignment (Fig. 3A). The MLST sequence-based phylogeny groups *N. meningitidis* genomes faithfully according to their sequence type (ST) and serogroup. For instance, all 3 ST-11 genomes (FAM18, M15141, and M9261) group together, as do the serogroup B genomes (M10699 and MC58). The most important feature of this tree is the fact that the disease-associated and carried isolates do not form separate and distinct monophyletic groups. In fact, disease-associated and carriage genomes are grouped together on this tree with high bootstrap support. This finding indicates that the excess of SNPs that discriminate between disease-associated and carried isolates is not based on shared evolutionary history alone.

In an attempt to gain more resolution for phylogenetic analysis, pairwise distances computed from the entire whole-genome alignment were used (Fig. 3B). This version of the phylogeny is slightly different with respect to some of the less supported internal branches, but disease-associated and carried genomes still do not form distinct mutually exclusive evolutionary groups. On this tree, several of the carried genomes represent basal evolutionary lineages that are nested between more derived lineages made up of disease-associated genomes that are closely related to each other but distantly related to other disease-associated isolates. The ST-198 carried isolate groups most closely with the serogroup B-invasive isolates on the whole-genome phylogeny, suggesting that it may have a highly chimeric genome sequence.

Distribution of SNP variation among *N. meningitidis* genomes. In addition to phylogenetic analysis, we directly evaluated how SNP variation is distributed within and between disease-associated and carried genomes using a population genetic measure, the fixation index (F_{ST}). F_{ST} is a population differentiation measure that is based on polymorphism data; it measures the difference of between-population variation from within-population variation. High values of F_{ST} (close to 1) indicate that polymorphisms tend to segregate between rather than within groups and reveal highly differentiated populations. Using the SNP variation data in the whole-genome sequence alignment, we measured F_{ST}, taking the disease-associated and carried groups of genomes as two putative populations (Fig. 4). F_{ST} for disease-associated versus carried genomes is low (0.03) and statistically indistinguishable from a null distribution of F_{ST} values calculated using a simulation

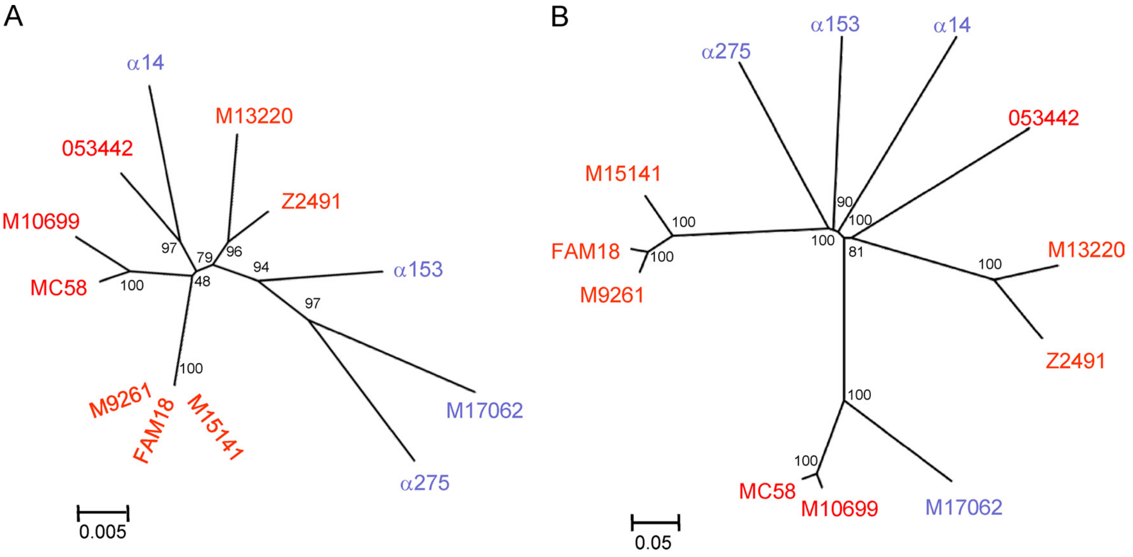

FIG. 3. Phylogenetic analysis of disease-associated and carried *N. meningitidis* isolate genomes. Disease-associated genomes are shown in red, and carried genomes are in blue. (A) A tree based on a concatenated alignment of the multilocus sequence typing loci: *abcZ*, *adk*, *aroE*, *fumC*, *gdh*, *pdhC*, and *pgm*. (B) A tree based on a whole-genome alignment.

procedure similar to that described for the discriminating SNP analysis ($z = 0.46$, $P < 0.32$). In other words, the polymorphism data based on the whole-genome alignment do not provide evidence for population subdivision between disease-associated and carried genomes of *N. meningitidis* as an explanation for the excess of observed discriminating SNPs.

We also used a naïve Bayesian classification approach to partition SNP variation among the *N. meningitidis* genomes studied here using *K*-means clustering (see Fig. S2 in the supplemental material). This approach was implemented with the program STRUCTURE in order to address two questions: (i) what is the optimal value of *K* (in other words, how many genome groups do the SNP data indicate) and (ii) are disease-associated and carried genomes segregated into distinct groups based on the SNP data? Based on a user-defined value of *K*, STRUCTURE assesses the statistical likelihood of observing the data given *K* and assigns individual SNPs into each group. For any given genome, the fraction of SNPs in each group can then be ascertained. This allows for a determination of the extent to which a given genome faithfully maps to one group or the other. The optimal value of *K* according to this analysis is 3, not 2, as may be expected if disease-associated and carried genomes formed distinct groups (see Fig. S2A in the supplemental material). In addition, the SNP variation at $K = 3$ does not cleanly partition among individual genomes (see Fig. S2B in the supplemental material). This result is consistent with a high level of recombination among *N. meningitidis* strains and is indicative of reticulate evolution and/or shared polymorphisms. This is particularly true for the majority of the carried isolate genomes, which appear to have the most mixed ancestry in terms of the three SNP clusters. Disease-associated genomes are less hybrid in general with respect to SNP polymorphism, and there are 7 disease-associated isolates with SNPs that segregate almost perfectly into 1 of 3 clusters. Apparently, there have been abundant opportunities for genetic exchange subsequent to the divergence of these genomes, and specific nucleotide variants acquired via recombination may be important markers for virulence.

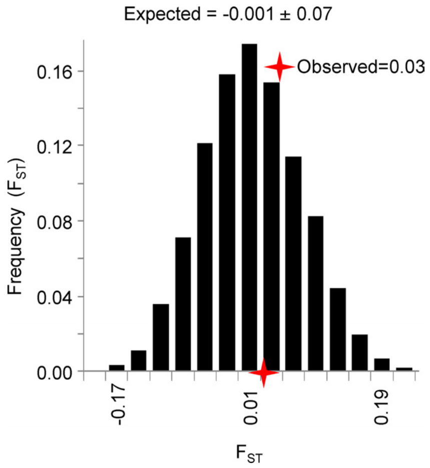

FIG. 4. Differentiation of SNPs within and between groups of disease-associated and carried *N. meningitidis* genomes based on the fixation index (F_{ST}). The *N. meningitidis* genome sequence alignment was simulated to yield a null distribution of F_{ST} values. The distribution of expected F_{ST} values is compared to the observed value (red cross).

Association of discriminating SNPs with *N. meningitidis* genes. Discriminating SNPs between disease-associated and carried genomes were associated with genes if they were found either within or proximal to coding sequences (Table 3). A total of 63 discriminating SNPs were associated with 45 *N. meningitidis* genes. Six of these discriminating SNPs map to gene-proximal noncoding sequences, and 57 map to coding sequences. The 57 coding sequence discriminating SNPs were classified as either nonsynonymous or synonymous based on whether or not they correspond to differences in encoded amino acid sequences between disease-associated and carried

TABLE 3. Genomic features and numbers of discriminating SNPs

Discriminating SNP class	Count
Noncoding gene associated	6
Coding sequence gene associated	57
Nonsynonymous coding sequence	40.75
Synonymous coding sequence	16.25

genomes. There are 40.75 (71.5%) nonsynonymous discriminating SNPs compared to only 16.25 (28.5%) synonymous SNPs. Below, we explore the possible functional implications of discriminating SNPs in more detail.

Functional characteristics of *N. meningitidis* discriminating SNP-associated genes. The SNP-associated genes (proteins) were evaluated with respect to a wide variety of functional characteristics to determine if they are enriched for any particular functions or features that may be related to virulence. The SNP-associated genes were not found to be enriched for function at the cell periphery (i.e., signal peptides or lipoproteins), horizontally transferred genes, or putative virulence factors. However, the SNP-associated genes were found to be enriched for 21 specific gene ontology (GO) functional annotations (Table 4). Many of the enriched SNP-associated genes span multiple functions across the GO hierarchy; there are a total of 21 SNP-associated genes among the overrepresented functional categories (Table 5). Of the 21 genes with overrepresented functions, we highlight two that are most closely related to virulence and explore the potential functional relevance of these SNP-associated genes below. These two SNP-associated genes represent a potential list of prioritized targets for future experimental interrogation based on the initial genome comparisons done here.

Although *mviN* is uncharacterized and is a putative gene in meningococcus, its products have been experimentally characterized in *Escherichia coli* and *Salmonella enterica* serovar Typhimurium, which are also Gram negative (3, 18). A mutation in the functional *S.* Typhimurium homolog of *mviN* renders an otherwise avirulent isolate virulent. In *E. coli*, *mviN* has been shown to be essential for murein synthesis.

NhhA has a few purported functions, including evading complement deposition and the resulting formation of the membrane attack complex (MAC) (44) as well as autotransporter and adhesin activity (41). Δ*nhhA* mutants have reduced adherence to host cells and have more MAC deposition. Therefore, NhhA is an adhesin, and it contributes to meningococcal immune evasion.

Dynamics of *N. meningitidis* colonization, carriage, and disease. The *N. meningitidis* genomes analyzed here were chosen based on the fact that the isolates originated from either diseased individuals, what we refer to as disease-associated isolates, or asymptomatic carriers, referred to here as carried isolates. The carried isolate genomes analyzed here are also distinguished by the fact that their STs are rarely or never associated with disease. Nevertheless, what we have observed in this analysis is essentially a snapshot in time and place of a set of particular nucleotide variants that distinguish one group of *N. meningitidis* disease-associated isolates from a group of carried isolates. Furthermore, it must be noted that in any individual, an invasive meningococcal disease case originates from an asymptomatic colonization state (19, 30). Transition of the bacterium from a colonization to disease-causing state may

TABLE 4. Overrepresented functions and categories in SNP-associated genes[a]

Category[b]	Description	No. of genes in the genome[c]	No. of SNP-associated genes[d]	*P* value[e]
GO:0009405	Pathogenesis	2	2	4.46e−4
GO:0044403	Symbiosis, encompassing mutualism through parasitism	2	2	4.46e−4
GO:0044419	Interspecies interaction between organisms	2	2	4.46e−4
GO:0016616	Oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	23	4	1.19e−3
GO:0050661	NADP or NADPH binding	11	3	1.33e−3
GO:0016614	Oxidoreductase activity, acting on the CH-OH group of donors	24	4	1.41e−3
GO:0051704	Multiorganism process	5	2	4.28e−3
GO:0016491	Oxidoreductase activity	147	8	1.09e−2
GO:0055086	Nucleobase, nucleoside, and nucleotide metabolic processes	66	5	1.20e−2
GO:0046483	Heterocycle metabolic process	105	6	2.21e−2
GO:0006753	Nucleoside phosphate metabolic process	53	4	2.48e−2
GO:0009117	Nucleotide metabolic process	53	4	2.48e−2
GO:0048037	Cofactor binding	83	5	2.97e−2
GO:0034641	Cellular nitrogen compound metabolic process	119	6	3.81e−2
GO:0006007	Glucose catabolic process	15	2	3.92e−2
GO:0019320	Hexose catabolic process	15	2	3.92e−2
GO:0046365	Monosaccharide catabolic process	15	2	3.92e−2
GO:0044275	Cellular carbohydrate catabolic process	16	2	4.42e−2
GO:0046164	Alcohol catabolic process	16	2	4.42e−2
GO:0006807	Nitrogen compound metabolic process	456	15	4.53e−2
GO:0044248	Cellular catabolic process	38	3	4.56e−2

[a] Categories of SNP-associated genes were analyzed using the hypergeometric distribution against the background of all genes.

[b] Gene ontology (GO) functional category.

[c] Number of genes in the M13220 reference genome belonging to this category (2,108 total genes in the genome).

[d] Number of SNP-associated genes belonging to this category (45 total SNP-associated genes).

[e] *P* value for the hypergeometric testing for enrichment of the GO category.

TABLE 5. SNP-associated genes from virulence-related overrepresented gene categories

Name[a]	Locus tags[b]	Functional annotation	Total no. of gene-associated SNPs	No. of SNPs flanking the coding sequence	No. of SNPs within the coding sequence		
					Total	Nonsynonymous	Synonymous
nhhA	NMA1200, NMB0992, NMC0978	Putative surface fibril protein	1	0	1	1	0
mviN	NMA2210, NMB0277, NMC0272	Putative inner membrane protein	3	3	0	0	0

[a] Consensus gene name of the homolog found in strains Z2491/MC58/FAM18.

[b] Locus tag of the homologous gene in strains Z2491/MC58/FAM18.

be accompanied by the acquisition of specific nucleotide variants; it may be dependent on the genetic background of the human host (11) or other environmental factors, or it may be caused by some combination of these factors. Thus, it is formally possible that the carried isolate genomes studied here could evolve rapidly to become invasive genomes that cause disease. Indeed, we have shown here that disease-associated and carried genomes do not form mutually exclusive evolutionary groups, consistent with repeated changes between these states over time (Fig. 3 and 4; see also Fig. S2 in the supplemental material). Even closer evolutionary relationships between disease-associated and carried isolates of *N. meningitidis* have been demonstrated elsewhere (19). The dynamics of all these factors necessitate that the discriminating SNPs and SNP-associated genes identified here be treated with caution, since they could change depending on changes in the disease-causing potential of the genomes in which they are found.

Conclusion. Because the presence or absence of genes alone does not determine meningococcal virulence (5, 16, 33, 42, 45, 46), we sought smaller differences in the form of SNPs between the genomes of 8 disease-associated and 4 carried isolates of *N. meningitidis*. We identified 63 discriminating SNPs and the genes associated with them. Of the 45 SNP-associated genes identified, functional analysis indicates 21 as the most likely targets of further investigation based on their possible roles in virulence.

In addition to the caveats described previously, it should be noted that the analyses performed here are limited by the relatively small number of complete genome sequences that were analyzed: 8 disease-associated and 4 carried isolates. Consequently, the particular set of discriminating SNPs characterized here, along with the list of SNP-associated genes, will likely change as additional genome sequences are characterized and compared. Thus, a more definitive understanding of genome-level differences between disease-associated and carriage isolates of *N. meningitidis* will require the analysis of additional genomes.

ACKNOWLEDGMENTS

This work was supported in part by the Centers for Disease Control and Prevention (1 R36 GD 000075-1) (L.S.K.), the Bioinformatics Program of the Georgia Institute of Technology (N.V.S.), the Georgia Research Alliance (GRA.VAC09.O) (B.H.H., I.K.J., and L.W.M.), and the Alfred P. Sloan Foundation (BR-4839) (I.K.J.).

We thank the Meningitis and Vaccine Preventable Diseases Branch at the CDC, especially Nancy Messonnier and Tom Clark, for helpful discussions. We also thank Ahsan Huda for helpful discussions and our anonymous reviewers, who gave us constructive feedback.

We also thank each organization or individual who provided some of the invasive isolates in this study. We thank the Active Bacterial Core Surveillance Team (M10699), the Emerging Infection Programs Network (M15141), the Research Institute for Tropical Medicine (M13220), Idrissa Sanou (M9261), and the collaboration between Ruth Lynfield, Billie A. Juni, Debbie Kirby, Henry Wu, and Cynthia P. Hatcher (M17062).

REFERENCES

1. **Altschul, S. F., et al.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:**3389–3402.
2. **Angellotti, M. C., S. B. Bhuiyan, G. Chen, and X. F. Wan.** 2007. CodonO: codon usage bias analysis within and across genomes. Nucleic Acids Res. **35:**W132–W136.
3. **Benjamin, W. H., Jr., J. Yother, P. Hall, and D. E. Briles.** 1991. The *Salmonella typhimurium* locus *mviA* regulates virulence in Itys but not Ityr mice: functional *mviA* results in avirulence; mutant (nonfunctional) *mviA* results in virulence. J. Exp. Med. **174:**1073–1083.
4. **Bentley, S. D., et al.** 2007. Meningococcal genetic variation mechanisms viewed through comparative analysis of serogroup C strain FAM18. PLoS Genet. **3:**e23.
5. **Bille, E., et al.** 2005. A chromosomally integrated bacteriophage in invasive meningococci. J. Exp. Med. **201:**1905–1913.
6. **Castillo-Davis, C. I., and D. L. Hartl.** 2003. GeneMerge—post-genomic analysis, data mining, and hypothesis testing. Bioinformatics **19:**891–892.
7. **Chen, L., et al.** 2005. VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res. **33:**D325–D328.
8. **Claus, H., M. C. Maiden, R. Maag, M. Frosch, and U. Vogel.** 2002. Many carried meningococci lack the genes required for capsule synthesis and transport. Microbiology **148:**1813–1819.
9. **Claus, H., et al.** 2005. Genetic analysis of meningococci carried by children and young adults. J. Infect. Dis. **191:**1263–1271.
10. **Darling, A. C., B. Mau, F. R. Blattner, and N. T. Perna.** 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. **14:**1394–1403.
11. **Davila, S., et al.** 2010. Genome-wide association study identifies variants in the CFH region associated with host susceptibility to meningococcal disease. Nat. Genet. **42:**772–776.
12. **Dolan-Livengood, J. M., Y. K. Miller, L. E. Martin, R. Urwin, and D. S. Stephens.** 2003. Genetic basis for nongroupable *Neisseria meningitidis*. J. Infect. Dis. **187:**1616–1628.
13. **Dull, P. M., et al.** 2005. Neisseria meningitidis serogroup W-135 carriage among US travelers to the 2001 Hajj. J. Infect. Dis. **191:**33–39.
14. **Emanuelsson, O., S. Brunak, G. von Heijne, and H. Nielsen.** 2007. Locating proteins in the cell using TargetP, SignalP and related tools. Nat. Protoc. **2:**953–971.
15. **Holmes, E. C., R. Urwin, and M. C. Maiden.** 1999. The influence of recombination on the population structure and evolution of the human pathogen *Neisseria meningitidis*. Mol. Biol. Evol. **16:**741–749.
16. **Hotopp, J. C., et al.** 2006. Comparative genomics of *Neisseria meningitidis*: core genome, islands of horizontal transfer and pathogen-specific genes. Microbiology **152:**3733–3749.
17. **Hudson, R. R., M. Slatkin, and W. P. Maddison.** 1992. Estimation of levels of gene flow from DNA sequence data. Genetics **132:**583–589.
18. **Inoue, A., et al.** 2008. Involvement of an essential gene, *mviN*, in murein synthesis in *Escherichia coli*. J. Bacteriol. **190:**7298–7301.
19. **Jolley, K. A., D. J. Wilson, P. Kriz, G. McVean, and M. C. Maiden.** 2005. The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in *Neisseria meningitidis*. Mol. Biol. Evol. **22:** 562–569.
20. **Joseph, B., et al.** 2010. Comparative genome biology of a serogroup B carriage and disease strain supports a polygenic nature of meningococcal virulence. J. Bacteriol. **192:**5363–5377.

21. **Juncker, A. S., et al.** 2003. Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Sci. **12:**1652–1662.
22. **Kingsford, C. L., K. Ayanbule, and S. L. Salzberg.** 2007. Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. Genome Biol. **8:**R22.
23. **Kislyuk, A. O., et al.** 2010. A computational genomics pipeline for prokaryotic sequencing projects. Bioinformatics **26:**1819–1826.
24. **Kumar, S., M. Nei, J. Dudley, and K. Tamura.** 2008. MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief. Bioinform. **9:**299–306.
25. **Lingappa, J. R., et al.** 2003. Serogroup W-135 meningococcal disease during the Hajj, 2000. Emerg. Infect. Dis. **9:**665–671.
26. **Maiden, M. C., et al.** 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc. Natl. Acad. Sci. U. S. A. **95:**3140–3145.
27. **Maiden, M. C. J., and D. A. Caugant.** 2006. The population biology of *Neisseria meningitidis*: implications for meningococcal disease, epidemiology and control, p. 17–35. *In* M. Frosch and M. C. J. Maiden (ed.), Handbook of meningococcal disease: infection biology, vaccination, clinical management. Wiley-VCH Verlag GmbH & Co., Weinheim, Germany.
28. **Margulies, M., et al.** 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature **437:**376–380.
29. **Mayer, L. W., et al.** 2002. Outbreak of W135 meningococcal disease in 2000: not emergence of a new W135 strain but clonal expansion within the electophoretic type-37 complex. J. Infect. Dis. **185:**1596–1605.
30. **Meyers, L. A., B. R. Levin, A. R. Richardson, and I. Stojiljkovic.** 2003. Epidemiology, hypermutation, within-host evolution and the virulence of *Neisseria meningitidis*. Proc. Biol. Sci. **270:**1667–1677.
31. **Parkhill, J., et al.** 2000. Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. Nature **404:**502–506.
32. **Peng, J., et al.** 2008. Characterization of ST-4821 complex, a unique *Neisseria meningitidis* clone. Genomics **91:**78–87.
33. **Perrin, A., et al.** 2002. Comparative genomics identifies the genetic islands that distinguish *Neisseria meningitidis*, the agent of cerebrospinal meningitis, from other *Neisseria* species. Infect. Immun. **70:**7063–7072.
34. **Pop, M., A. Phillippy, A. L. Delcher, and S. L. Salzberg.** 2004. Comparative genome assembly. Brief. Bioinform. **5:**237–248.
35. **Pritchard, J. K., M. Stephens, and P. Donnelly.** 2000. Inference of population structure using multilocus genotype data. Genetics **155:**945–959.
36. **Pruitt, K. D., T. Tatusova, and D. R. Maglott.** 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. **35:**D61–D65.
37. **Rangannan, V., and M. Bansal.** 2010. High-quality annotation of promoter regions for 913 bacterial genomes. Bioinformatics **26:**3043–3050.
38. **Rosenstein, N. E., B. A. Perkins, D. S. Stephens, T. Popovic, and J. M. Hughes.** 2001. Meningococcal disease. N. Engl. J. Med. **344:**1378–1388.
39. **Rusniok, C., et al.** 2009. NeMeSys: a biological resource for narrowing the gap between sequence and function in the human pathogen *Neisseria meningitidis*. Genome Biol. **10:**R110.
40. **Saitou, N., and M. Nei.** 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:**406–425.
41. **Scarselli, M., et al.** 2006. *Neisseria meningitidis* NhhA is a multifunctional trimeric autotransporter adhesin. Mol. Microbiol. **61:**631–644.
42. **Schoen, C., et al.** 2008. Whole-genome comparison of disease and carriage strains provides insights into virulence evolution in *Neisseria meningitidis*. Proc. Natl. Acad. Sci. U. S. A. **105:**3473–3478.
43. Reference deleted.
44. **Sjolinder, H., J. Eriksson, L. Maudsdotter, H. Aro, and A. B. Jonsson.** 2008. Meningococcal outer membrane protein NhhA is essential for colonization and disease by preventing phagocytosis and complement attack. Infect. Immun. **76:**5412–5420.
45. **Snyder, L. A., and N. J. Saunders.** 2006. The majority of genes in the pathogenic *Neisseria species* are present in non-pathogenic *Neisseria lactamica*, including those designated 'virulence genes.' BMC Genomics **7:**128.
46. **Stabler, R. A., et al.** 2005. Identification of pathogen-specific genes through microarray analysis of pathogenic and commensal *Neisseria* species. Microbiology **151:**2907–2922.
47. **Stein, L. D., et al.** 2002. The generic genome browser: a building block for a model organism system database. Genome Res. **12:**1599–1610.
48. **Tettelin, H., et al.** 2000. Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. Science **287:**1809–1815.
49. **Thompson, J. D., T. J. Gibson, and D. G. Higgins.** 2002. Multiple sequence alignment using ClustalW and ClustalX. Curr. Protoc. Bioinform. **2:**2.3.
50. **Vernikos, G. S., and J. Parkhill.** 2006. Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. Bioinformatics **22:**2196–2203.
51. **Yang, J., L. Chen, L. Sun, J. Yu, and Q. Jin.** 2008. VFDB 2008 release: an enhanced Web-based resource for comparative pathogenomics. Nucleic Acids Res. **36:**D539–D542.
52. **Zdobnov, E. M., and R. Apweiler.** 2001. InterProScan—an integration platform for the signature-recognition methods in InterPro. Bioinformatics **17:**847–848.