Supporting Information

Buckee *et al.* 10.1073/pnas.0712019105

SI Text

Microepidemic Model Simulations. Simulations were performed in which bacterial isolates, defined by seven loci, were repeatedly sampled from a population assuming the microepidemic clustering process modeled by Fraser *et al.* (1). Using comparable rates of mutation and recombination, and a skewed sampling framework based on a Poisson distribution (1), mismatch distributions were measured over time. As in the Fraser *et al.* model, a higher number of isolates had zero mismatches than expected because of the clusters of strains caused by the microepidemic process. Over time, however, although the distribution of mismatches remained stable, the composition of the clusters with respect to the alleles at each of the seven loci changed rapidly. Fig. S1 shows the results of 1,000 simulations, with the number of identical strains present in the sample as time progresses measured on the *y* axis. With repeated sampling, the chances of finding the same strain over more than a few time steps rapidly decreases. Most STs in the Czech dataset are short-lived, whereas others such as ST44 are recovered throughout the 27 years of surveillance. Fig. 1 shows that these two observations cannot be reconciled within the same neutral microepidemic framework.

Expected Lifespan of Lineages. In our data, 64% of isolates collected each year represented "new" combinations of housekeeping genes on average. The expected lifespan of housekeeping gene combinations was modeled by using this rate of appearance, with a hypothetical bacterial population being sampled over many years. Fig. S2 shows the cumulative appearance of housekeeping gene allele combinations observed in the data (in blue) and a histogram of simulated lifespans (yellow). None of the lineages in the simulations existed for >7 years.

Additional Results for the Deterministic Model of Competition Between Lineages. Fig. S3 shows the range of model (as defined by Eqs. **1** and **2**) outcomes for a hypothetical five-locus, two-allele system under different levels of competition for available hosts. Competition for hosts is not specifically attached to any particular mechanism in this model and thus could as easily arise from factors other than immune selection such as extended colonization or antibiotic use. Under low levels of competition, all possible genotypes (each designated here by a binary bar code) can coexist and acquire excess virulence, the relative abundance of these variants being dependent upon their respective transmissibilities (Fig. S3*A*). As competition increases, the genotypes of lower transmissibility decline in frequency, as generally also observed within multistrain model with cross-immunity, and are no longer able to carry excess virulence (as this has the effect of further reducing their transmissibility), as shown in Fig. S3*B*. With a further increase in competition, some of these genotypes drop out altogether (Fig. S3*C*), and gradually the population is dominated by only a few of the many possible combinations of housekeeping genes, each constituting a distinct stable lineage (Fig. S3*D*), with excess virulence emerging only among the more transmissible lineages who are able to tolerate this penalty. Under very intense competition, the population is dominated by a very limited number of genotypes, none of which possess excess virulence (Fig. S3*E*).

Stochastic Model of Immune Selection with Lineage Structure. Strains were defined by two antigenic determinants, to which hosts gained allele-specific immunity, as well as an ST, which determined the transmissibility of the strain but had no effect on immunity. For simplicity, each locus had two alleles: one antigenic loci expressing *a* or *b*, and the other *x* or *y*, and the strain types were simply labeled 1 or 2. Thus, there were eight possible strains, *ax*1, *ay*1, *bx*1, *by*1, *ax*2, *ay*2, *bx*1, and *by*2. Two strains coinfecting the same host could recombine at a given rate, which was the same for each locus and allowed genes to respond to different types of selection pressure. Mutation also occurred at the same rate for each locus, facilitating the emergence of new variants. Each potential host was a separate entity including information about its contacts, the strains it was infected with, and its immune response (memory of infection). Contacts between hosts were random and changed at every time step, and changes in host infection and immunity status are updated synchronously at the end of each time step. Upon infection, hosts remained infectious for a period (such that the probability of losing infectiousness at each time step was $1/\sigma$, where σ is the average duration of infectiousness), and then retained immunity to the antigenic determinants for a further period (again, the probability of losing immunity at each time step was $1/\varsigma$, where ς was the average duration of immunity). Hosts became completely susceptible again once immunity was lost.

Cross-immunity was incorporated such that infection by one strain conferred partial protection against other strains expressing the same antigenic determinants, regardless of their strain type. Competition between strain types was generated by differences in their R_0 values, and it was assumed that ecological constraints did not allow for the coinfection of the same host by two strains belonging to the same strain type. Upon infection, individuals remained infected by that pathogen for a period, such that the average duration of infection with a pathogen was $1/m$, where m is the probability that the host rids itself of the pathogen in a time step. After infection, individuals remained immune to that pathogen for a period, such that the average duration of immunity to that pathogen was $1/s$, where s is the probability of the host losing its immunity to a pathogen in a time step. The duration of infection and immunity therefore exhibit exponential decay. When an infection event occurs there is also the chance that the strain will undergo mutation or recombine with another strain in the same host. Both of these events occur with defined probabilities (t and r, respectively). The ranges of the parameters explored were chosen to encompass what is known of meningococcal biology (2), and are outlined in Table S4. Parameter space was explored by using the statistical technique of Latin Hypercube Sampling (LHS) (3), which selects combinations of parameter values without replacement, given parameter value ranges and probability distribution functions. The key model parameters that were sampled using LHS can be found in Table S4. We used several thousand LHS samples to cover parameter space.

Antigenic determinants and strain types followed distinctly different trajectories. The effect of cross-immunity on the population structuring of just the antigenic determinants was similar to previous models (4–6), with weak cross-immunity allowing for all variants to coexist, and increasing levels resulting in a rapid transition to a population dominated by two stable, nonoverlapping "antigenic islands," for example, variants *ax* and *by*. The other variants, in this case *ay* and *bx*, were suppressed, even though they were continuously generated by recombination, because most hosts were immune to dominant strains sharing these epitopes.

The behavior of the STs, which defined only the transmissibility of the strain, depended on the competition between them

and the structuring of the antigenic determinants. When cross-immunity was low and all of the antigenic variants were in circulation, the STs coexisted within each combination of antigens according to their transmissibility. Ecological competition between strain types caused them to oscillate within these islands, however their overall prevalence was reflective of their R_0 values. Thus, all eight strains cocirculated, unless the fitness difference between the two STs was large enough to cause competitive exclusion, in which case one ST dominated each antigenic variant. Here, the frequency with which strain types oscillated within the antigenic island depended on the rates of recombination and mutation, with increased rates causing more rapid oscillations ($R^2 = 0.18$ and 0.23, respectively; $P < 0.0001$).

When cross-immunity was high, however, generating stable antigenic islands of nonoverlapping epitopes, the dynamics of the STs depended on the differences between their R_0 values. For these discretely structured populations, the STs followed one of three distinct trajectories within the two dominant antigenic islands. The particular outcome depended primarily on the level of competition between the two STs, as defined by the difference between their R_0 values. Fig. 2 shows the median and range of these differences for the three groups for all simulations; all three were significantly different from one another. To confirm the impact of competition on the dynamics of the STs further simulations were run with high levels of cross-immunity, in which the transmissibilities of the two STs were equal for 3,000 time steps, and then changed slightly for another 3,000 time steps. As expected, initially the STs were stably associated with one antigenic variant each. However, once the transmissibilities had been altered slightly, oscillations occurred at a frequency dependent on the fitness differences between them. The frequency with which strain types oscillated with respect to their antigenic variants, for those simulations where the antigenic types were discrete, also depended mainly on competition. Larger differences between the transmissibilities of the STs led to increased frequency of oscillation ($R^2 = 0.25$; $P < 0.0001$), unlike the populations in which the antigenic determinants were not structured, where recombination and mutation rates were the significant parameters. Although population size affected both the time taken to reach a steady state and the probability of extinction, it did not otherwise affect the dynamics of the model. We have also explored a deterministic analogue of this model that shows that these outcomes can be generalized to multiple discrete antigenic and strain types.

1. Fraser C, Hanage WP, Spratt BG (2005) Neutral microepidemic evolution of bacterial pathogens. *Proc Natl Acad Sci USA* 102:1968–1973.
2. Frosch M, Maiden MCJ (2006) *Handbook of Meningococcal Disease: Infection Biology, Vaccination, Clinical Management* (Wiley, New York).
3. Blower SM, Dowlatabadi H (1994) Sensitivity and uncertainty analysis of complex models of disease transmission: An HIV model, as an example. *Int Stat Rev* 2:229–243.
4. Gupta S, *et al.* (1996) The maintenance of strain structure in populations of recombining infectious agents. *Nat Med* 2:437–442.
5. Gupta S, Ferguson N, Anderson RM (1998) Chaos, persistence, and evolution of strain structure in antigenically diverse infectious agents. *Science* 280:912–915.
6. Buckee CO, Koelle K, Mustard MJ, Gupta S (2004) The effects of host contact network structure on pathogen diversity and strain structure. *Proc Natl Acad Sci USA* 101:10839–10844.

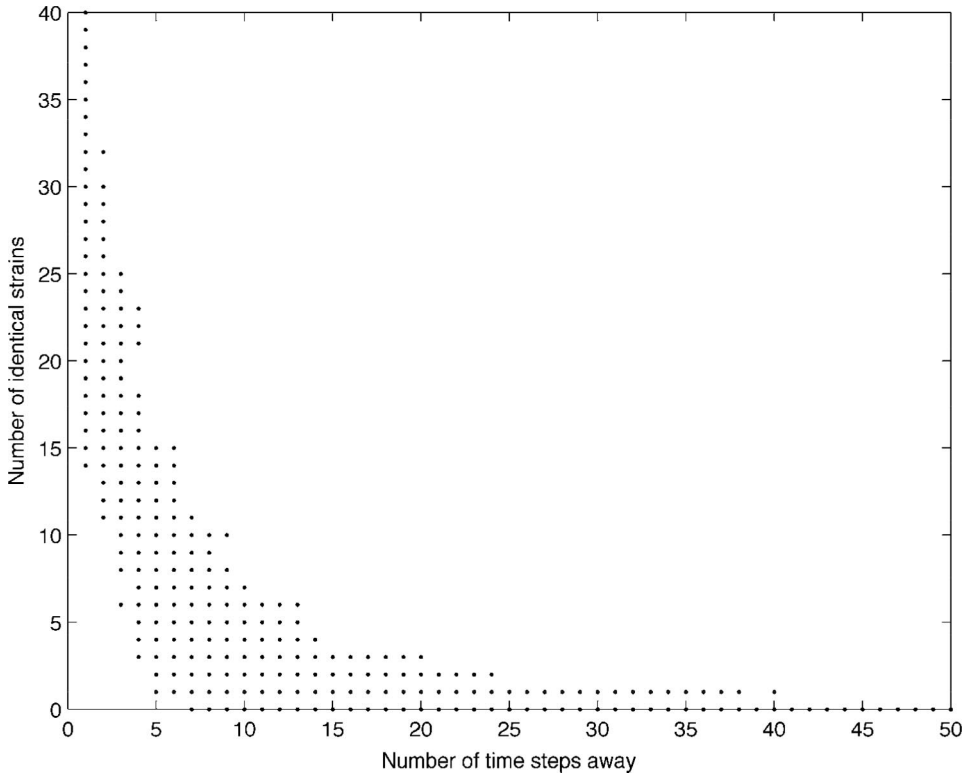

Fig. S1. The decay of allele combinations over time assuming a microepidemic process as in Fraser *et al.* (1). For each strain, the number of identical strains found in the sample at different numbers of time steps away is shown. Shown are the results of 1,000 simulations of repeated sampling, and each data point represents multiple results.

1. Fraser C, Hanage WP, Spratt BG (2005) Neutral microepidemic evolution of bacterial pathogens. *Proc Natl Acad Sci USA* 102:1968–1973.

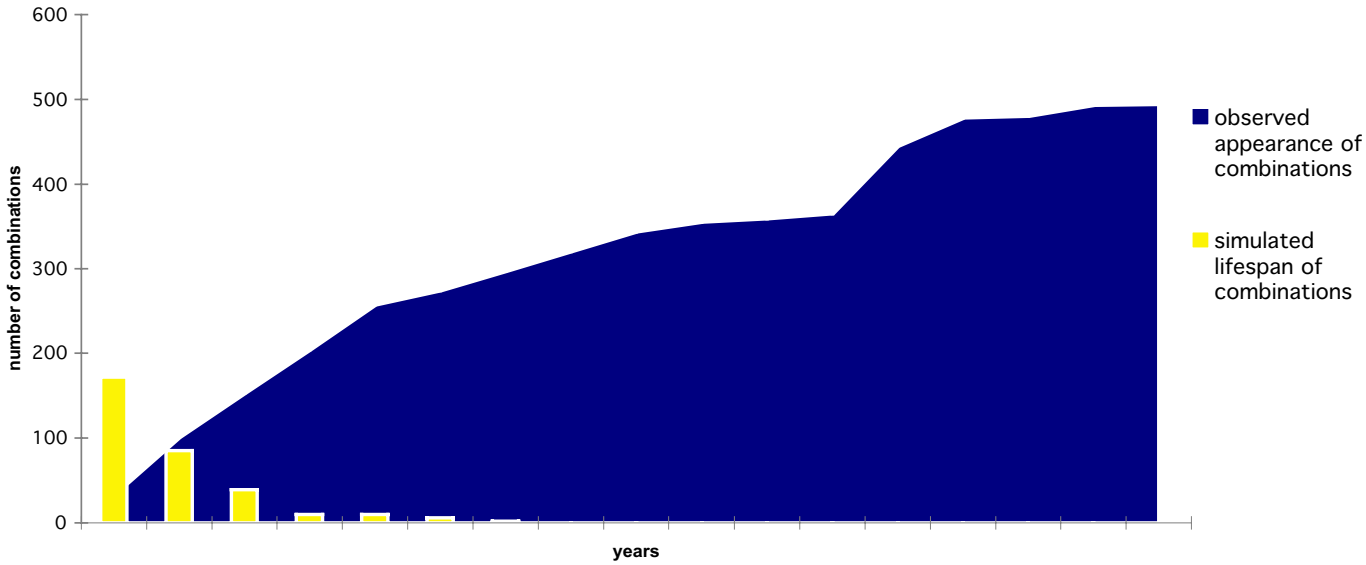

Fig. S2. The cumulative appearance of lineages observed in the data (in blue) compared with the simulated lifespan of lineages (in yellow). Assuming the rate of appearance shown, the majority of lineages persist for only 1 year, and none persist for >7 years.

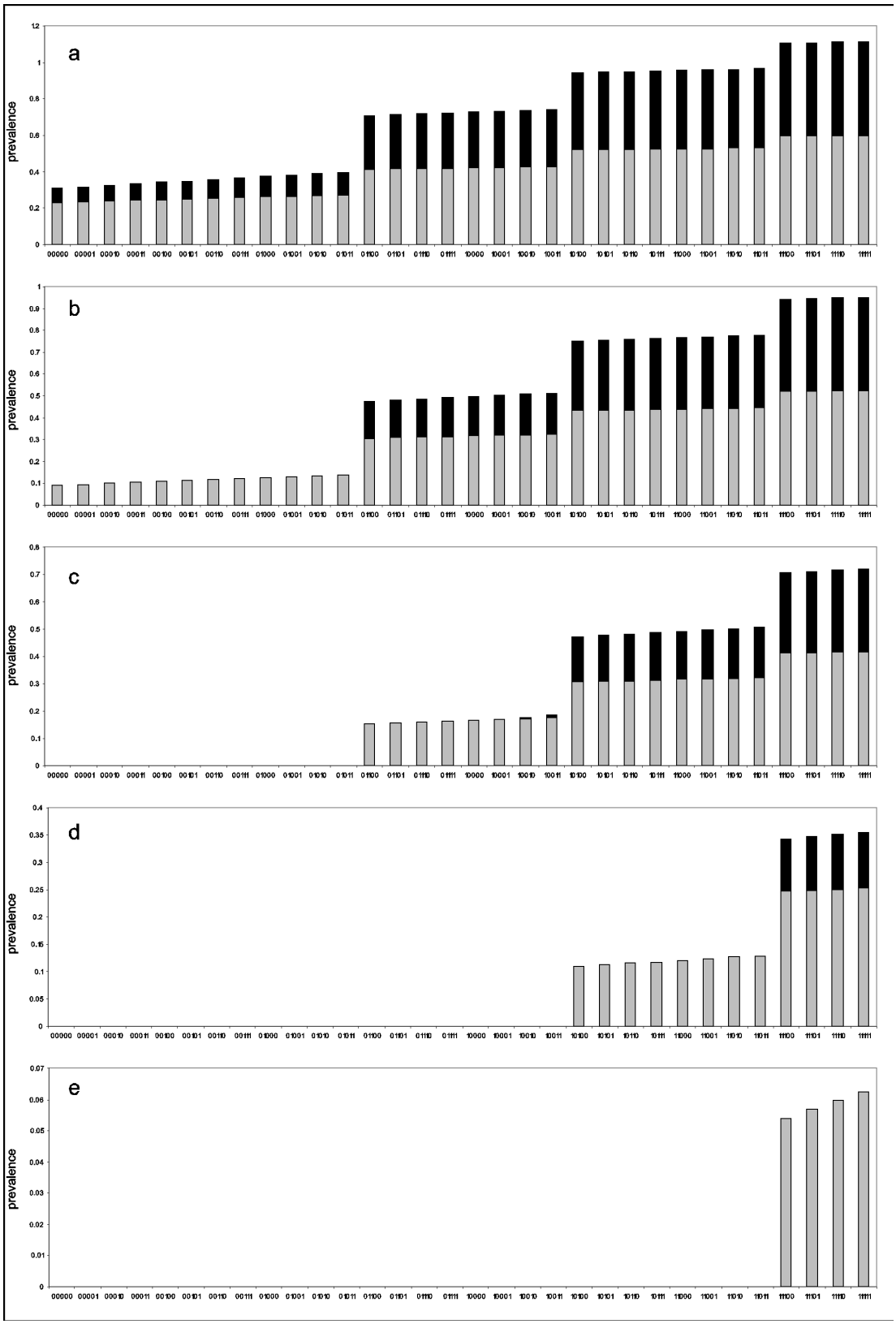

Fig. S3. The prevalences at equilibrium of avirulent (gray) and virulent (black) members of 32 genotypes (defined by combinations of housekeeping genes within a hypothetical five-locus, two-allele system and each designated by a binary bar code) exhibiting an ascending order of transmission coefficients $\beta_i = \beta + 0.1 \cdot i$, with $\beta = 20$ for $1 < i \leq 12$, $\beta = 25$ for $12 < i \leq 20$, $\beta = 30$ for $20 < i \leq 28$, and $\beta = 35$ for $28 < i \leq 32$. Levels of competition: $\gamma = 0.35$ (*a*), $\gamma = 0.45$ (*b*), $\gamma = 0.55$ (*c*), $\gamma = 0.65$ (*d*), and $\gamma = 0.75$ (*e*). Other parameter values: $\mu =$ o.o2/yr, $\sigma = 10$/yr, $\alpha_0 = 0$, $\alpha_1 = 2$.

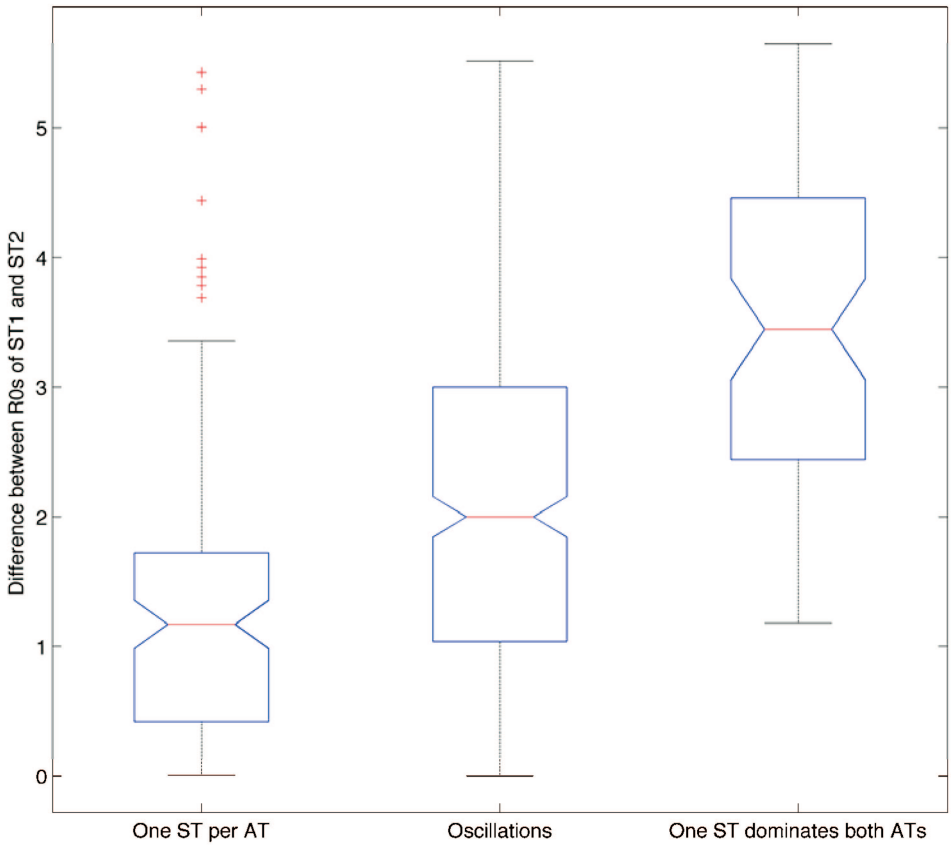

Fig. S4. The effects of competition, defined as the difference between the transmissibilities of the two STs, on the dynamics of their association with discrete antigenic islands under high cross-immunity. The box plot shows the median, lower, and upper quartiles and the extent of the range of competitive values for all simulations, separated into the different dynamics. Notches show a robust estimate of uncertainty about the median, and nonoverlapping notches between groups indicate that the medians are significantly different at the 5% level. The significance of the comparison of these groups by ANOVA is $F = 63.3$; $P < 0.000001$.

Table S1. Observed frequencies of clonal complexes per year

	Year											
Clonal complex	1971	1972	1973	1974	1975	1978	1979	1980	1993	1994	1996	Total
cc41/44	5	26	11	9	25	8	4	8	31	7	3	137
cc549	1	34	14	2	4	1	1		5	2		64
cc376	7	11	5	17	6	2	2	3		1	1	55
cc22	2	2	5	11	10	2	3	5	3			43
cc92	2		1	3		4			21	5	6	42
cc11									33	4		37
cc174	2	9	6	2	7	1			1	3		31
cc254		3	3	9	5	3	1		1	1	3	29
cc106									20	4	3	27
cc226	1	1		2	2	4	6	7		2		25
cc269	1	3	11		1	3	1	3	1			24
cc116							1		13	8		22
cc292	3	4	1	5	2		2	1	1	2		21
cc53									12	6	2	20
cc364	5	1	2	3	3	1	2					17
cc37		2		1	2			8		2		15
cc231			3	1	1		2	2	3	2		14
cc18				3					5	3		11
cc865		1		6	1					1		9
cc32									1	1	5	7
cc103									4		1	5
cc1	1	1		1				1				4
cc750		1	1	1				1				4
cc334	1		1						1			3
cc23				1		1						2
cc35									1		1	2
cc167				1						1		2
Unassigned	19	36	31	24	46	16	19	21	61	24	8	305
Total	50	135	95	102	115	46	44	60	218	79	33	977

Table S2. Pairwise F_{ST} values between years for concatenated MLST allele sequences

	1971	1972	1973	1974	1975	1978	1979	1980	1993	1994	1996
1971	0										
1972	−0.00566	0									
1973	0.00003	0.00595	0								
1974	0.005	0.01066*	0.03214*	0							
1975	−0.00479	−0.0045	0.002	0.00755	0						
1978	−0.00279	−0.00969	0.00303	0.02034*	−0.0037	0					
1979	−0.00557	0.00154	0.00849	0.02979*	0.00725	−0.00608	0				
1980	−0.00084	0.0037	0.00743	0.00143	−0.00031	0.00684	0.00866	0			
1993	0.02416*	0.01947*	0.01637*	0.05247*	0.02322*	0.00888	0.02554*	0.03177*	0		
1994	0.01423	0.01392*	0.00527	0.0474*	0.01065	0.00719	0.02302*	0.02425*	−0.00423	0	
1996	0.00581	0.00799	0.01058	0.04371*	0.01895	0.0005	0.01522	0.02749*	−0.00532	−0.00199	0

*$P < 0.05$.

Table S3. Clonal complex association of PorA variants (%) in the 1970s and 1990s

	PorA VR1:VR2 variants																									
	18–1,3		5–1,2–2		19,15		5–1,10–1		5,2		21,16		18,25		22,14–6		12–1,13–1		5–2,10		7–1,1		5–1,10–4		*N*	
Complex	70 s	90 s	70 s	90 s	70 s	90 s	70 s	90 s	70 s	90 s	70 s	90 s	70 s	90 s	70 s	90 s	70 s	90s	70 s	90 s	70 s	90 s	70 s	90s	70 s	90 s
cc41/44				25.0	2.0	59.1					13.3	3.7	89.3		77.8	50.0	19.0				13.6			16.7	89	44
cc549			69.8									7.4			3.7				25.0				7.7		57	7
cc376	2.0				72.5	4.5							3.6		7.4						9.1				52	2
cc92							50.0	45.2			6.7	3.7											7.7	41.7	11	32
cc22	31.0																		18.8	50.0			3.8		38	3
cc11		14.3							100																	37
cc174	25.0				2.0																		3.8		27	4
cc254			4.8		2.0						46.7	11.1									9.1		3.8		24	5
cc106								52.4				3.7														27
cc269			1.6															20.0	6.3						22	1
cc116												48.1													1	21
cc292			4.8		3.9						13.3	3.7								25.0	4.5		3.8		18	3
cc53												3.7						40.0								21
cc226	13.0	28.6	1.6																12.5						17	2
cc364	1.0																47.6		6.3						17	
cc231	4.0																	20.0			13.6				8	5
cc18		14.3	4.8																			33.3		8.3	3	8
cc865	1.0		7.9		2.0														6.3						8	1
cc37																					22.7	66.7			5	2
cc103																				25.0						6
cc334		14.3															4.8								3	1
cc1																			18.8						3	
cc750					2.0																				3	
cc167																							3.8	8.3	1	1

Values are the percentages of each variant associated with each clonal complex in the decade specified.

Table S4. Parameter ranges explored during simulations of the stochastic model

Parameter	Description	LHS range, min/max
C	Mean number of contacts per host	4:20
$1/\mu$	Average duration of infection	3:10
$1/\sigma$	Average duration of immunity	10:30
β	Probability of transmission (to a completely susceptible host)	0.2:0.8
R	Probability of recombination per allele	0.01:0.1
τ	Probability of allelic mutation per allele	0.001:0.005
γ	Degree of cross-immunity	0.02:4
π	Population size	200:1,000