Neutral microepidemic evolution of bacterial pathogens

Christophe Fraser*†, William P. Hanage*, and Brian G. Spratt

Department of Infectious Disease Epidemiology, St. Mary's Hospital Campus, Imperial College London, Norfolk Place, London W2 1PG, United Kingdom

Edited by John J. Mekalanos, Harvard Medical School, Boston, MA, and approved December 27, 2004 (received for review September 21, 2004)

Understanding bacterial population genetics is vital for interpreting the response of bacterial populations to selection pressures such as antibiotic treatment or vaccines targeted at only a subset of strains. The evolution of transmissible bacteria occurs by mutation and localized recombination and is influenced by epidemiological as well as molecular processes. We demonstrate that the observed population genetic structure of three important human pathogens, *Streptococcus pneumoniae*, *Neisseria meningitidis*, and *Staphylococcus aureus*, can be explained by using a simple evolutionary model that is based on neutral mutational drift, modulated by recombination, and which incorporates the impact of epidemic transmission in local populations. The predictions of this neutral ''microepidemic'' model are found to closely fit observed genetic relatedness distributions of bacteria sampled from their natural population, and it provides estimates of the relative rate of recombination that agree well with empirical estimates. The analysis suggests the emergence of neutral bacterial population structure from overlapping microepidemics within clustered host populations and provides insight into the nature and size distribution of these clusters. These findings challenge the assumption that strains of bacterial pathogens differ markedly in relative fitness.

infinite-alleles model | multilocus sequence typing | recombination

It is now accepted that bacteria do not conform to the clonal model of evolution (1). The importance of recombination has become increasingly clear in recent years, both as a fundamental process in strain diversification (2) and as a mechanism by which strains acquire virulence factors or resistance determinants (3). Homologous recombination in bacteria involves the replacement of a small segment of the bacterial chromosome (a few kilobases) with the corresponding region from another isolate (2). The frequency of these localized recombinational events may be extremely rare, resulting in species that are highly clonal [e.g., *Mycobacterium* species (4, 5)], or extremely frequent, resulting in species that are almost completely nonclonal [e.g., *Helicobacter pylori* (6)]. Consequently, theoretical approaches developed for exclusively sexual or asexual organisms are inappropriate, and, at present, there is no general theory reconciling these variable properties of bacteria. The problem is further complicated by serious sampling biases arising as a result of overrepresentation of disease isolates or antibiotic-resistant isolates in clinical strain collections (1). This problem is especially acute for some species, including *Streptococcus pneumoniae*, *Staphylococcus aureus*, and *Neisseria meningitidis*, which we focus on here and which are ''accidental'' pathogens in that healthy carriage is common, with disease a rare outcome. Finally, the host population structure will influence transmission and needs to be accounted for when considering bacterial population structure: Spread of directly transmitted bacteria within a social group is much more likely than between randomly chosen hosts from the population as a whole (7, 8).

Methods

We develop a multilocus model of bacterial evolution that incorporates varying levels of recombination (Fig. 1). To fit the predictions of the model to empirical data requires representative samples of the natural population. Here we use four samples from cross-sectional studies of carriage within a local area (9–12) in which isolates were characterized by using multilocus sequence typing (MLST), a technique in which DNA sequences are obtained for seven housekeeping loci and the different sequences at each locus are assigned as different alleles (13). The samples are described in more detail in Table 3, which is published as supporting information on the PNAS web site.

Results

A null model for evolutionary change is the neutral infinite-alleles model (IAM) (14), in which mutation and drift are the primary determinants of gene frequencies. We extend the IAM model to include variation at multiple loci and varying levels of localized recombination (Fig. 1). To compare the predictions of our evolutionary model with observed population genetic structure, we initially use the distribution of pairwise allelic mismatches: i.e., the proportion of pairs of isolates that differ at zero, one, two, or more of the seven sequenced loci (Fig. 2, filled bars). This distribution has been used previously to detect linkage between loci and to infer the degree of clonality within different species (1, 15), because the expected allelic mismatch distribution can easily be computed in the limit of complete linkage equilibrium (16). Interestingly, this distribution is remarkably similar for the two samples of *S. pneumoniae* from infants at different locations (Fig. 2*A*), despite marked differences in the strains composing each sample [only 17% are present in both samples, and those that are have markedly different frequencies in each (Fig. 3*A*)]. We derive an analytical expression for this distribution in our model (see *Appendix*); the best fit is shown in Fig. 2 as a dotted line.

This model, which has only two parameters, namely, the rate at which new alleles enter the population θ and the rate at which they are shuffled by recombination ρ, superficially fits this distribution quite well for all three species. Simulations show, however, that the model consistently differs significantly at the leftmost bar (Fig. 2, open circles), corresponding to an underestimate of the frequency of pairs of isolates that are identical at all loci. The model also is inconsistent with other features of the data. Specifically, it overestimates the number of different strains in the sample (Table 2), does not reproduce the ''nearest neighbor'' distribution (Fig. 3*B* and Fig. 5, which is published as supporting information on the PNAS web site) and does not match the genotypic clustering as assessed by EBURST (17) (Table 2 and Table 4, which is published as supporting information on the PNAS web site). The purely neutral model can therefore be rejected. The deviation from the basic model is largely due to an excess of identical pairs of isolates in the natural populations. We initially account for this excess by introducing an empirical parameter, h_e, equivalent to the magnitude of this deviation (see

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: IAM, infinite-alleles model; MLST, multilocus sequence typing.

*C.F. and W.P.H. contributed equally to this work.

†To whom correspondence should be addressed. E-mail: c.fraser@imperial.ac.uk.

www.pnas.org/cgi/doi/10.1073/pnas.0406993102

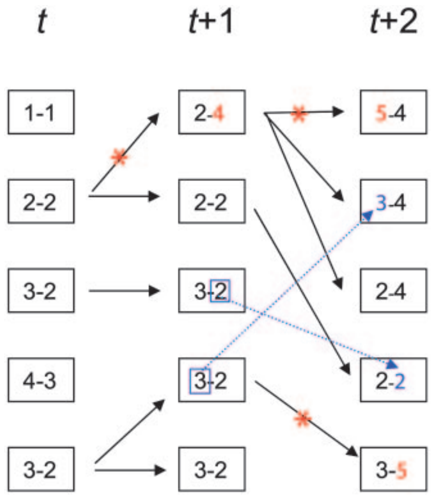


Fig. 1. A neutral multilocus infinite alleles model of bacterial evolution. Schematic illustrating the model for a population of five individuals. The bacterial strain infecting each individual is characterized by two integers that identify the alleles at two loci. At time *t,* for example, there are two cases of colonization by bacteria of genotype 3-2. At each generation, each individual can infect any other (represented by black arrows). Mutations occur during the transmission step with rate *m* and are indicated by red asterisks: Each mutation always generates a new allele. Recombination events, occurring with rate *r* and illustrated by blue dotted arrows, result in an allele being inherited from a random donor. Mutations and recombination events can affect more than one allele in a single step (not shown) and are not exclusive. More generally, the model is defined for *i* loci in a population of size *N*. The model is simulated by starting from a single genotype until equilibrium levels of diversity are reached.

Appendix for formulas). In all four carriage samples, h_e is positive (Table 2) and significantly improved the fit to the data (solid lines in Fig. 2). We then explored possible mechanisms by which the excess of identical strains measured by h_e could be generated. The most obvious source is sampling bias (e.g., overrepresentation of isolates associated with disease or antibiotic resistance), which is unlikely because the populations studied in this work were specifically designed to minimize this type of bias by cross-sectional sampling of the natural carried population.

Another potential source of this excess is an inflation of the frequencies of certain strains through selective advantage. However, simulations of populations under selection result in negative estimates of the parameter h_e (Fig. 4*A*). This counterintuitive result does not indicate that selection reduces the proportion of identical pairs of isolates. Rather, selection alters the whole allelic mismatch distribution. Instead of only changing the leftmost bar, the proportions of strains that differ at multiple loci are also altered. If we naively fit a neutral (i.e., mis-specified) model to data generated with selection, these differences across the mismatch distribution lead to the best fit being a negative value of h_e.

A further potential source of deviation from the purely neutral model is infectious transmission: Two or more isolates may be identical because they form part of the same short transmission chain, which is likely in samples taken from a local population. Such "microepidemics" have been directly observed in families, daycare centers, and villages (7, 8). We therefore simulated neutral microepidemic evolution by using maximum likelihood estimated parameters (Table 1) and incorporating a final step to simulate epidemic linkage as measured by h_e (see *Appendix*). The consistency of the simulated populations with both the results of the analytical solution and real data are remarkable (results shown as filled circles in Figs. 2 and 3). The simulated populations were also analyzed by using EBURST (17), and the results are shown compared with real data in Fig. 3 *C* and *D*; summary statistics of EBURST output were found to be strikingly similar to observed patterns for all three species (Tables 2 and 4).

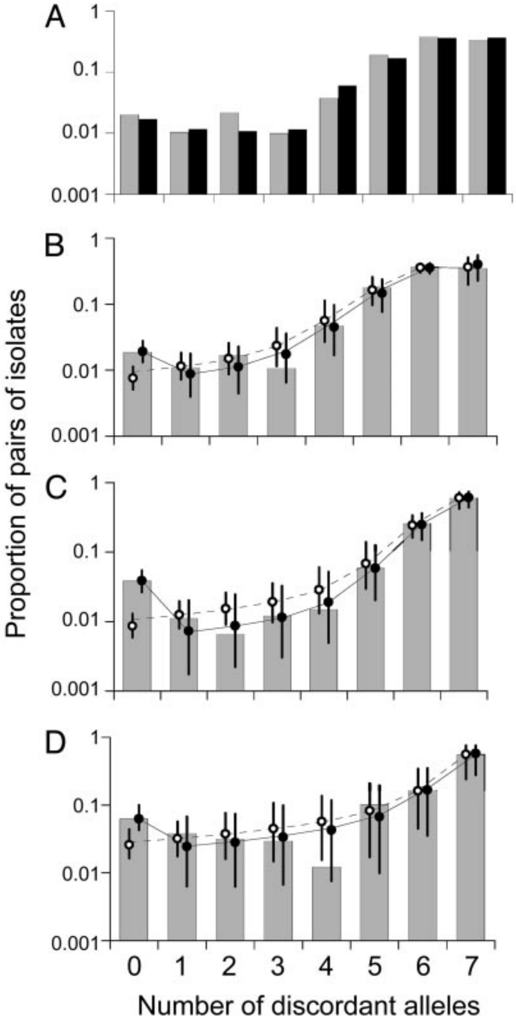


Fig. 2. Model fit to data. The allelic mismatch distributions F_k^7 are shown as filled bars for *S. pneumoniae* (*A* and *B*), *N. meningitidis* (*C*), and *Staphylococcus aureus* (*D*). (*A*) In the case of *S. pneumoniae*, two samples were included (Oxford, gray bars; Tampere, Finland, black bars). (*B*) Weighted mean of the two samples in *A*. The predictions of the purely neutral model and the neutral microepidemic model are shown fitted to samples as dashed and solid lines, respectively. Maximum likelihood parameter estimates are shown in Table 1. Parameter estimates obtained by fitting to the two pneumococcal studies independently were virtually identical to the joint estimate, reflecting the strong similarity in population structure seen in *A*. Simulation results are also shown (open circles for the neutral IAM and solid circles for the neutral microepidemic model), along with 95% prediction intervals.

The final sampling step in the model effectively reconstructs a local population, consisting of a limited number of sociospatial clusters, skewed in size, taken from a purely neutral global population. This relation between local and global populations arises dynamically, provided microepidemics are indeed restricted in size (Fig. 6, which is published as supporting information on the PNAS web site). We thus predict that for an unbiased global sample, very little epidemic linkage should be observed (i.e., $h_e \approx 0$); in the absence of such an unbiased sample, we examined the largest available sample of *S. pneumoniae*, the MLST database of isolates submitted from global sources ($n = 1{,}856$ at the time of study). Although these are not systematically sampled and, therefore, the results should be

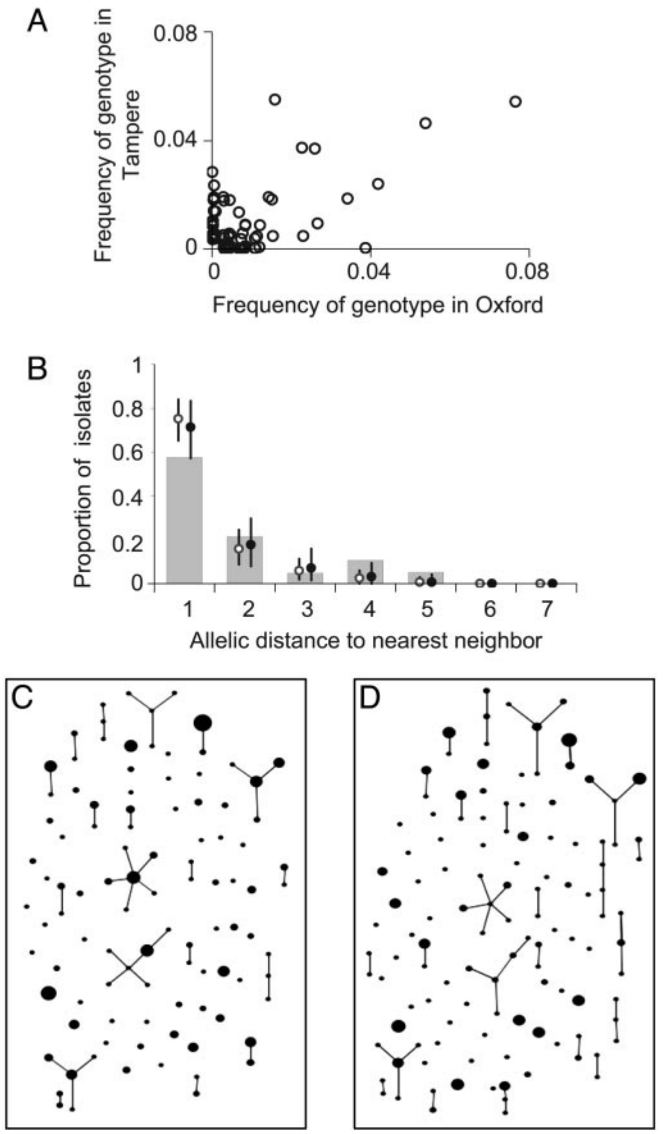


Fig. 3. Additional tests of the model. (*A*) To display differences in content between the two *S. pneumoniae* samples, we plotted the frequency of each genotype (strain) in each sample. (*B*) The agreement between simulation and data from the *S. pneumoniae* sample in Oxford was then examined by using the nearest neighbor distribution, defined as the proportion of isolates whose distance to the most similar nonidentical isolate is $k = 1...7$, shown plotted as a function of k. Results from the purely neutral model are shown as open circles, and the neutral microepidemic model results are shown as solid circles. The same distribution was plotted for the three other samples in Fig. 6. Differences in fit between the models are slight, reflecting the fact that the greatest difference is an excess of homozygosity, to which the nearest neighbor analysis is relatively insensitive. We also performed EBURST analysis of the Oxford data set (*C*) and a single realization of the neutral microepidemic model (*D*) simulated by using parameters from Table 1. Each different strain is represented by a point, the size of which is the frequency of the strain. Strains differing at a single locus, which are inferred to be linked by descent, are joined by lines. A summary of the clustering inferred by EBURST is shown in Table 2.

treated with caution, the estimate of h_e is very close to zero (0.002). We also predict from this model that if we were to combine different local samples from a global population, only the estimate of h_e would change. Analysis of the combined pneumococcal data sets supports this view, because neither θ nor ρ changes; as expected, the estimate of h_e is reduced (0.0087). This value is not as low as predicted by complete independence of the samples (0.0059), but we believe the question of whether this difference is caused by some epidemiological linkage between the populations, chance, or other factors can only be resolved by gathering further samples.

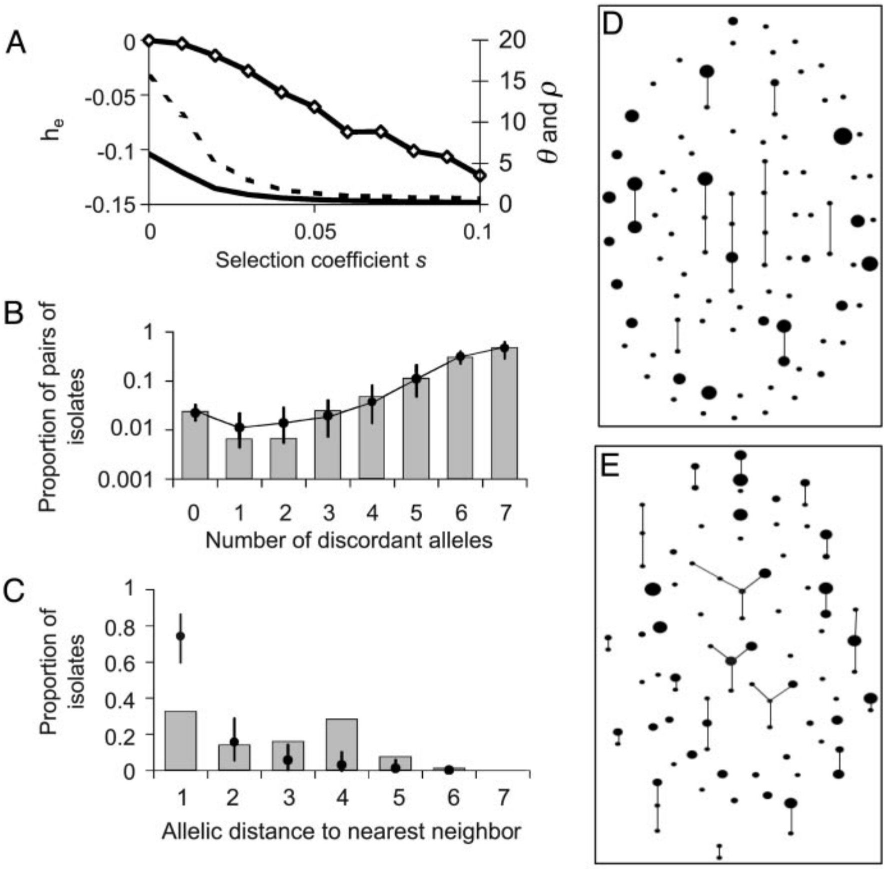
The principal effect of microepidemic population structure at larger scales is to reduce the effective population size to a fraction of the number of infected hosts (Fig. 6). More realistic nonlinear scaling of mean cluster size with census size would reduce the effective population size still further. The suggestion that the basic unit of transmission is in some sense larger than the individual infected host is in accordance with recent developments in theoretical epidemiology, which have redefined key parameters such as the basic reproduction number, R_0, for macroepidemics in terms of transmission between closely linked clusters of individuals rather than between individuals themselves (18).

Although methods are available that estimate the bacterial recombination rate from sequence data, this remains a major computational challenge (19, 20). Our model can estimate this quantity with ease from multilocus allelic data (e.g., MLST data). To test the validity of the estimates, we calculate the ratio ($\rho/\theta = r/m$) and compare it with empirical estimates obtained by a modification of the method of Feil *et al.* (21) (Table 1); these two estimates are essentially independent, because the approach of Feil *et al.* examines only the most recent evolutionary changes (those generating strains with differences at single loci; second bar from left in Fig. 2), whereas our estimate uses the entire mismatch distribution but gains most of its information from distantly related pairs of strains because these are far more frequent. This concordance offers further support for our underlying model, and we note that this estimate of r/m is robust to variation in h_e [unlike previous methods based on the index of association (1)]. The neutral microepidemic model also estimates the extent of epidemiological clustering in real data (h_e), which allows estimation of two new parameters: the number of clusters, n_c, and their mean size, $\bar{\sigma}$. We found quite different values for these parameters for the three species (Table 1), suggesting differences in transmission patterns. Interestingly, *S. pneumoniae*, which is not typically associated with outbreaks, had the lowest mean cluster size ($\bar{\sigma}$). In contrast, *N. meningitidis*, which is more associated with community outbreaks, had a larger mean cluster size.

It is often assumed that strains of bacterial pathogens differ markedly in fitness, and it is surprising that, after accounting for microepidemics, the observed population structures fit a neutral model. We therefore attempted to fit the neutral microepidemic model to samples generated from a simulation incorporating selection. Although the fit to the allelic mismatch distribution was acceptable (Fig. 4*B*), the model comprehensively failed to capture other features of these populations (Fig. 4 *C*–*F*). Thus, if selection had played a major role in structuring the bacterial populations we examined, we would have expected a poor fit to these metrics. Preliminary analyses of a diverse range of scenarios, including direct selection, balancing selection, population subdivision, hypermutation, and hyperrecombination, all failed to generate results consistent with the data. We cannot, however, exclude the possibility that much more complex models could fit the data as well as, or better than, that which we propose here. Nonetheless, we are struck by the success of this simple model, although we recognize the need to further test it against such alternative hypotheses.

Discussion

We have developed a model of bacterial evolution and tested it with samples from three different species. This model is defined by only three parameters: the population mutation and recombination rates and the degree of epidemic linkage in the sample. The model successfully captures the observed structure, mea-



F

	N$_{Strains}$	N$_{eBurst\ Groups}$	N$_{singletons}$
Simulated data with selection	91	9	68
Fitted Neutral Micro-epidemic Model (95% PI)	86.3 (77.5-95.1)	17.2 (13.2-21.2)	39.8 (32.2-47.4)

Fig. 4. Analysis of a model with hitch-hiking selection. Selection acts upon a single unobserved locus that entirely determines the fitness of a strain. Variation occurs at the same population diversification rates, θ and ρ, as the MLST-defining loci. Mutation causes the fitness of an allele to be multiplied by a log-normal random deviate of mean 1 and standard deviation *s*, the selection coefficient. We also tested a normal random deviate, and truncated distributions including only beneficial or harmful mutations. Results were similar in each case. (*A*) The allelic mismatch distribution obtained for the neutral microepidemic model is refitted to simulated populations with no clustering, produced with $\theta = 5.7$, $\rho = 17.1$, and varying values of the selection coefficient. The fit remains good, but selection results in reduced estimates of θ (solid line) and ρ (dashed line) and negative values of h_e (diamonds). (*B*–*F*) A single sample is drawn from a simulated population with both selection and clustering, with $\theta = \rho = 64$, $N = 2{,}000$, and $s = 0.1$. The sample of size $n = 250$ includes $n_c = 25$ epidemic clusters of mean size $\bar{\sigma} = 6$, and we attempt to refit the neutral microepidemic model to this sample produced with selection. (*B*) The allelic mismatch distribution fits acceptably, resulting in estimates $\theta = 6.4$, $\rho = 9.6$, and $h_e = 0.015$. However, the resulting nearest neighbor distribution (*C*) and the EBURST analyses (*D*, sample simulated with selection and clustering; *E*, sample from best-fit neutral microepidemic model; *F*, summary statistics) fit poorly.

sured by using multiple metrics, of the four samples studied (Figs. 2, 3, and 6 and Tables 2 and 4). The flexibility of our model and the ease of computing the key parameters should make it ideal for further exploring the effect of different epidemic scenarios on the population genetics of a species and for realistic parameterized simulations of evolutionary scenarios. We have also shown that differences in transmission patterns may be detected by using this approach.

A key finding is that given the well known phenomenon of microepidemics within host clusters, the population structure of the three pathogens studied is consistent with neutral drift. The poor fit of samples generated under selection to the model support our view that the imprint of selection is not present in our four population samples, although it would be interesting to explore our ability to detect selection by using samples where selection should be present, such as those exposed to a new vaccine or antibiotic. These findings challenge us to either identify the signature of selection by other means or to accept that the common assumption that directly transmitted pathogens must be subject to strong selection is not supported by the data. This latter conclusion has implications for modeling and public health.

Appendix

Analytic Expression for the Allelic Mismatch Distribution. Consider the neutral multilocus IAM with recombination, described in Fig. 1, and define the allelic mismatch distribution $F_k^i(t)$ as the probability that any two isolates differ at *k* of *i* studied loci at time *t*. The aim is to obtain equilibrium expressions by considering

Table 1. Parameter estimates

	Neutral model		Neutral microepidemic model			Epidemic clusters		Relative recombination rate		
Species	θ	ρ	θ	ρ	h_e	n_c	$\bar{\sigma}$	n_{rec}:n_{mos}:n_{mut}	$(r/m)_{\mathrm{pred}}$	$(r/m)_{\mathrm{obs}}$
S. pneumoniae	5.0	12.4	5.3	17.3	0.011	22/24	5.0/5.8	44:6:15	2.7	2.1
N. meningitidis	8.2	5.7	10.2	13.6	0.033	9	13.1	13:7:5	1.2	1.1
Staphylococcus aureus	4.6	0.37	5.6	0.98	0.026	7	13.6	2:0:19	0.13	0.11

θ and ρ are, respectively, the maximum likelihood population mutation and recombination rates obtained for the neutral multilocus model. The model fit is improved significantly by the introduction of the parameter h_e, which allows for an excess of identical pairs of isolates. n_c and $\bar{\sigma}$ are the number and mean size of the clusters inferred from the samples (the two values for *S. pneumoniae* are for the Tampere, Finland, and Oxford studies, respectively). n_{rec}, n_{mos} and n_{mut} are the number of pairs of isolates differing at a single locus that are classified as being the result of whole-locus recombination, mosaic recombination (i.e., recombination between a donor and recipient that occurs within the allele and produces a new mosaic allele), and point mutation by adapting the empirical method of Feil *et al.* (21). Patterns of descent among closely related genotypes were determined by using EBURST (17) to identify ancestral and descendant alleles among strains differing at a single locus. Recombination was identified if the descendant allele was found in other lineages (EBURST groups) within the sample (22). Variants differing at a single base pair were assigned as mutations; the remainder were identified as mosaic recombination. $(r/m)_{\mathrm{obs}}$ is the resulting empirical estimate of the relative recombination rate, i.e., $(r/m)_{\mathrm{obs}} = n_{\mathrm{rec}}/(n_{\mathrm{mos}} + n_{\mathrm{mut}})$; $(r/m)_{\mathrm{pred}}$ is the value predicted by our model adjusted for homozygous recombination, i.e., $(r/m)_{\mathrm{pred}} = F_1'^1\rho/\theta$.

first the changes that occur during a single generation. For a single locus, the distribution is unaffected by recombination, and thus the classic result of Kimura (14), $F_0^1 = 1/(1 + \theta)$, holds, where m is the per locus mutation rate, N is the population size, and $\theta = 2mN$. For more loci, consider first the probability F_0^i that a pair of isolates is identical at all loci. We study the model in the limit where it is vanishingly unlikely that two or more events could occur simultaneously, although, in fact, the result can be shown numerically to be valid even away from this limit. In a generation, there are three events that could affect this: the isolates could be from an identical progenitor (with probability $1/N$), in which case they are always identical; one of the isolates could mutate (with total probability $2im$), in which case they will fail to be identical; or they could recombine (with total probability $2ir$), where r is the per locus recombination rate. The effect of recombination is to separate the inheritance at the recombinant locus from the others, thus reducing the pair comparison to that between the recombinant locus and the $i - 1$ others. In summary, the change is

$$F_0^i(t + 1) = (1 - 1/N - 2im - 2ir)F_0^i(t) + 1/N + 2irF_0^{i-1}(t)F_0^1(t), \qquad [1]$$

which results in the equilibrium expression

$$F_0^i = \frac{1 + i\rho F_0^{i-1}F_0^1}{1 + i\theta + i\rho}, \qquad [2]$$

where we have defined the population recombination rate $\rho = 2rN$ by analogy with θ. For the more general expression F_k^i, where k > 0, note that the mismatch will increase to F_{k+1}^i if a mutation occurs at any of the $i - k$ identical loci, but it can be reached from F_{k-1}^i if mutation occurs at any of the $i - k + 1$ identical loci. In the case of recombination, the possibility that the recombinant locus may be either concordant or discordant must be accounted for. The change in a single generation is thus

$$F_k^i(t + 1) = (1 - 1/N - 2(i - k)m - 2ir)F_k^i(t) + 2(i - k + 1)mF_{k-1}^i + 2ir(F_{k-1}^{i-1}(t)F_1^1(t) + F_k^{i-1}(t)F_0^1(t)), \qquad [3]$$

which results in the equilibrium expression

$$F_k^i = \frac{(i - k + 1)\theta F_{k-1}^i + i\rho(F_k^{i-1}F_0^1 + F_{k-1}^{i-1}F_1^1)}{1 + (i - k)\theta + i\rho}. \qquad [4]$$

Table 2. Summary of EBURST analysis

	N_{Strains}	$N_{\mathrm{EBURST\ groups}}$	$N_{\mathrm{Singletons}}$
S. pneumoniae (Oxford study)	100	19	46
Neutral (95% PI)	145 (138.0–152.6)	30.1 (22.5–37.7)	40.3 (23.5–57.1)
Neutral microepidemic (95% PI)	97 (81.4–112.6)	18.5 (12.8–24.2)	45.2 (34.1–56.3)

Simulations were conducted by using the pure neutral and neutral microepidemic models to generate populations that were then analyzed by using EBURST. The numbers of strains, clusters of related genotypes (EBURST groups or clonal complexes), and genotypes that were distantly related to all others (singletons) are shown, as well as 95% prediction intervals (PI). Default settings for EBURST were used (17).

Fitting the Model. The model was fitted by maximizing the multinomial log-likelihood with respect to the parameters θ and ρ, which is given by

$$l(\theta, \rho) = \frac{n(n - 1)}{2}\sum_{k=0}^{i}\hat{F}_k^i \ln(F_k^i(\theta, \rho)), \qquad [5]$$

where n is the sample size, $\hat{F}_k^i$ is the observed allelic mismatch distribution, and additive constants have been ignored.

Modified Allelic Mismatch Distribution. A modified allelic mismatch distribution F'^i_k is introduced to allow for an excess of identical pairs of isolates by introducing the empirical parameter h_e as follows: $F'^i_0 = h_e + (1 - h_e)F_0^i$ and $F'^i_k = (1 - h_e)F_k^i$ for $k > 0$. The pure IAM model is recovered by setting $h_e = 0$. The likelihood remains as defined above but is now maximized with respect to the three parameters θ, ρ, and h_e. Improvement in fit was assessed by the likelihood ratio test, allowing for the extra parameter. Because the multinomial likelihood (Eq. **5**) overestimates the degrees of freedom in the data, we used the conservative replacement of $n(n - 1)/2$ by n in Eq. **5**. P values for the improved fit were 0.03 for Fig. 2*B*, <0.001 for Fig. 2*C*, and <0.01 for Fig. 2*D*.

Simulation of Epidemic Linkage in a Local Sample. Initially, we construct a truly neutral global sample. To constitute a locally clustered sample of size n, we took n_c samples from the global population in which a single isolate was included σ times (drawn from a Poisson distribution with mean $\bar{\sigma}$), and we completed the sample by taking randomly drawn isolates, included once. This skewed sampling process creates an excess of identical pairs relative to the underlying neutral population. The best fit values of the parameters n_c and $\bar{\sigma}$ are determined by matching to the number of distinct strains recorded in the sample, subject to the constraint that $h_e = n_c\bar{\sigma}^2/(n(n - 1))$, where the metaparameter h_e is determined by fitting the analytical formula to the allelic mismatch distribution. The formula h_e can be derived as follows. First, consider a cluster of size σ. This results in an extra $\sigma(\sigma - 1)/2$ identical pairs of isolates. The expected increase in identical pairs of isolates per cluster is $\bar{\sigma}^2/2$, and there are n_c such clusters. Thus, the proportionate increase in the number of identical pairs of the total $n(n - 1)/2$ is $h_e = n_c\bar{\sigma}^2/(n(n - 1))$.

We dedicate this work to the memory of John Maynard Smith. We thank Ed Feil and Roy Anderson for useful comments. This work was funded in part by The Medical Research Council (C.F.), The Wellcome Trust (W.P.H. and B.G.S.), and the Royal Society (C.F.).

1. Maynard Smith, J., Smith, N. H., O'Rourke, M. & Spratt, B. G. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 4384–4388.
2. Spratt, B. G., Hanage, W. P. & Feil, E. J. (2001) *Curr. Opin. Microbiol.* **4,** 602–606.
3. Ochman, H., Lawrence, J. G. & Groisman, E. A. (2000) *Nature* **405,** 299–304.
4. Supply, P., Warren, R. M., Banuls, A. L., Lesjean, S., Van Der Spuy, G. D., Lewis, L. A., Tibayrenc, M., Van Helden, P. D. & Locht, C. (2003) *Mol. Microbiol.* **47,** 529–538.
5. Smith, N. H., Dale, J., Inwald, J., Palmer, S., Gordon, S. V., Hewinson, R. G. & Smith, J. M. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 15271–15275.
6. Falush, D., Kraft, C., Taylor, N. S., Correa, P., Fox, J. G., Achtman, M. & Suerbaum, S. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 15056–15061.
7. Leino, T., Auranen, K., Jokinen, J., Leinonen, M., Tervonen, P. & Takala, A. K. (2001) *Pediatr. Infect. Dis. J.* **20,** 1022–1027.
8. Hope Simpson, R. E. (1952) *Lancet*, **260,** 549–554.
9. Jolley, K. A., Kalmusova, J., Feil, E. J., Gupta, S., Musilek, M., Kriz, P. & Maiden, M. C. (2000) *J. Clin. Microbiol.* **38,** 4492–4498, and correction (2002) **40** 3549–3550.
10. Hanage, W. P., Auranen, K., Syrjanen, R., Herva, E., Makela, P. H., Kilpi, T. & Spratt, B. G. (2004) *Infect. Immun.* **72,** 76–81.
11. Day, N. P., Moore, C. E., Enright, M. C., Berendt, A. R., Smith, J. M., Murphy, M. F., Peacock, S. J., Spratt, B. G. & Feil, E. J. (2001) *Science* **292,** 114–116, and retraction (2002) **295,** 971.
12. Meats, E., Brueggemann, A. B., Enright, M. C., Sleeman, K., Griffiths, D. T., Crook, D. W. & Spratt, B. G. (2003) *J. Clin. Microbiol.* **41,** 386–392.
13. Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D. A., *et al.* (1998) *Proc. Natl. Acad. Sci. USA* **95,** 3140–3145.
14. Kimura, M. (1968) *Nature* **217,** 624–626.
15. Whittam, T. S., Ochman, H. & Selander, R. K. (1983) *Proc. Natl. Acad. Sci. USA*, **80,** 1751–1755.
16. Brown, A. H. D., Feldman, M. W. & Nevo, E. (1980) *Genetics*, **96,** 523–536.
17. Feil, E. J., Li, B. C., Aanensen, D. M., Hanage, W. P. & Spratt, B. G. (2004) *J. Bacteriol.* **186,** 1518–1530.
18. Ball, F. & Neal, P. (2002) *Math. Biosci.* **180,** 73–102.
19. Stumpf, M. P. & McVean, G. A. (2003) *Nat. Rev. Genet.* **4,** 959–968.
20. McVean, G., Awadalla, P. & Fearnhead, P. (2002) *Genetics* **160,** 1231–1241.
21. Feil, E. J., Maiden, M. C., Achtman, M. & Spratt, B. G. (1999) *Mol. Biol. Evol.* **16,** 1496–1502.