Paper:

Numminen, E., Gutmann, M., Shubin, M., Marttinen, P., Méric, G., van Schaik, W., Coque, T., Baquero, F., Willems, R., Sheppard, S., Feil, E., Hanage, W. & Corander, J. (2016). The impact of host metapopulation structure on the population genetics of colonizing bacteria. *Journal of Theoretical Biology, 396*, 53-62.

http://dx.doi.org/10.1016/j.jtbi.2016.02.019

www.elsevier.com/locate/yjtbi

The impact of host metapopulation structure on the population genetics of colonizing bacteria

Elina Numminen, Michael Gutmann, Mikhail Shubin, Pekka Marttinen, Guillaume Méric, Willem van Schaik, Teresa M. Coque, Fernando Baquero, Rob J.L. Willems, Samuel K. Sheppard, Edward J. Feil, William P. Hanage, Jukka Corander

PII: S0022-5193(16)00115-6
DOI: http://dx.doi.org/10.1016/j.jtbi.2016.02.019
Reference: YJTBI8548

To appear in: *Journal of Theoretical Biology*

Received date: 7 August 2015
Revised date: 15 January 2016
Accepted date: 13 February 2016

Cite this article as: Elina Numminen, Michael Gutmann, Mikhail Shubin, Pekk Marttinen, Guillaume Méric, Willem van Schaik, Teresa M. Coque, Fernand Baquero, Rob J.L. Willems, Samuel K. Sheppard, Edward J. Feil, William P Hanage and Jukka Corander, The impact of host metapopulation structure on th population genetics of colonizing bacteria, *Journal of Theoretical Biolog* http://dx.doi.org/10.1016/j.jtbi.2016.02.019

The impact of host metapopulation structure on the population genetics of colonizing bacteria

Elina Numminen[1], Michael Gutmann[1,2], Mikhail Shubin[1], Pekka Marttinen[2], Guillaume Méric[3], Willem van Schaik[4], Teresa M. Coque[5], Fernando Baquero[5], Rob J. L. Willems[4], Samuel K. Sheppard[3], Edward J. Feil[6], William P. Hanage[7], Jukka Corander[1]

[1]Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland; [2]Helsinki Institute for Information Technology HIIT, Department of Information and Computer Science, Aalto University, Finland; [3]College of Medicine, Swansea University, Institute of Life Science, Swansea, UK; [4]Department of Medical Microbiology, University Medical Center Utrecht, Utrecht, The Netherlands; [5]Department of Microbiology, Ramón y Cajal University Hospital, Madrid, Spain; [6]Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath, UK; [7]Center for Communicable Disease Dynamics, Harvard School of Public Health, Boston, Massachusetts, USA;

Abstract

Many key bacterial pathogens are frequently carried asymptomatically, and the emergence and spread of these opportunistic pathogens can be driven, or mitigated, via demographic changes within the host population. These inter-host transmission dynamics combine with basic evolutionary parameters such as rates of mutation and recombination, population size and selection, to shape the genetic diversity within bacterial populations. Whilst many studies have focused on how molecular processes underpin bacterial population structure, the impact of host migration and the connectivity of the local populations has received far less attention. A stochastic neutral model incorporating heightened local transmission has been previously shown to fit closely with genetic data for several bacterial species. However, this model did not incorporate transmission limiting population stratification, nor the possibility of migration of strains between subpopulations, which we address here by presenting an extended model. We study the consequences of migration in terms of shared genetic variation and show by simulation that the previously used summary statistic, the allelic mismatch distribution, can be insensitive to even large changes in microepidemic and migration rates. Using likelihood-free inference with genotype network topological summaries we fit a simpler model to commensal and hospital samples from the common nosocomial pathogens *Staphylococcus aureus*, *Staphylococcus epidermidis*, *Enterococcus faecalis* and *Enterococcus faecium*. Only the hospital data for *E. faecium* display clearly marked deviations from the model predictions which may be attributable to its adaptation to the hospital environment.

Key words: Bacterial evolution, genetic structure, migration, population dynamics

Introduction

Bacteria colonizing multicellular hosts are organized in a hierarchy of local interconnected subpopulations forming a complex metapopulation as a whole. The subpopulations can range in scale from discrete intracellular colonies residing within a single host cell to pervasive strains circulating among hosts across cities, countries and continents (Fraser et al., 2009). Although most bacteria are harmless or even advantageous to their host organisms, some cause infectious disease, and understanding the evolutionary dynamics and the factors producing the genetic variation of pathogen populations is important for combatting disease emergence and spread.

Previous work has demonstrated that a simple model of stochastic microepidemics arising from repeated sampling of localized transmission chains, can explain genotypic variation in local surveillance data from several common human pathogens (Fraser et al., 2005; Hanage et al., 2006), under an assumption that all isolates are equally fit (neutrality). Conceptually, the microepidemic model is similar to the Λ-coalescent models (Pitman, 1999; Sagitov, 1999) which allow coalescence of multiple lineages simultaneously. However, under the microepidemic model no single genealogy would typically adequately capture the relatedness of samples since recombination is allowed to shuffle alleles across the haploid population. In the studies of Fraser et al. and Hanage et al., populations were characterized by a simple measure of the level of genotype relatedness known as the allelic mismatch distribution, which is defined by the empirical distribution of the allelic distance between pairs of individuals. Similar pairwise comparisons have been widely used in classical ecology and population genetics and different patterns in the mismatch distribution can be associated with various factors contributing to the population structure, including: population growth (Harpending, 1994; Rogers and Harpending, 1992), selection (Bamshad et al., 2002), and host contact network structure (Plucinski et al., 2011). The mismatch distribution has also been used to detect deviations from neutrality or constant population size (Mousset et al., 2004) and for inference about bacterial recombination rates (Hudson, 1987).

Population structure is one of the most studied phenomena in population genetics, both from the theoretical and applied perspective (Ewens, 2004; Hartl and Clark, 2007; Rousset, 2004). Both in bacteria and eukaryotes there are numerous situations where the effects of population structure are complex and challenging to quantify analytically. One of the main reason for this is the need to simultaneously account for several major forces known to impact the neutral evolution of populations. For bacterial pathogen populations it is typical to consider mutation, recombination and clonal expansion, whereas migration has been given less attention. Migration may for example be caused by anthroponosis and zoonosis when multiple different host organisms are colonized by the same bacterial species. The presence of multiple simultaneous forces hampers both the theoretical derivation of limit results for such models and empirical fitting due to likelihood equations not necessarily being available in closed form. Fraser et al. solved the likelihood intractability arising from microepidemics by using a stochastic mixture distribution to account for the increase in the probability of sampling identical strains from the same transmission chain (Fraser et al., 2005). An analogous approximation technique has later been independently introduced in a more general ecological setting and it is known as the *synthetic likelihood* (Wood, 2010).

To improve understanding of the evolutionary dynamics of structured bacterial populations, we employ a simulation-based approach to neutral models that can account for the multiple stochastic forces impacting the genetic diversity that persists over time. By capturing both a heterogeneous span of microepidemics and migration events across the boundaries limiting transmission between subpopulations, we characterize the expected behavior of the metapopulations as a whole. This provides an opportunity to explore the limits of inferring the model parameters from genetic surveillance data.

Materials and Methods

Model

We consider a neutral Wright-Fisher infinite alleles model for a finite haploid population with N individuals and discrete non-overlapping generations, where the reproduction takes place by random sampling of N individuals from the current generation to the next generation (Ewens, 2004). When the population is assumed structured, the subpopulation sizes are indexed by N_1, N_2. The parameters which may vary across subpopulations are indexed accordingly. We note that the host population which the haploid organism uses as a habitat remains implicit in our study and we consider no within-host variation. This would correspond to assuming a strong transmission bottleneck for each host in a standard epidemic model. Mutations are introduced per generation by a Poisson process with the rate $\theta = \mu N \tau$, where μ is the per locus mutation rate and τ is a scaling factor representing the generation time in calendar time. In all subsequent work we set $\tau = 1$, unless otherwise mentioned.

We assume that each individual is characterized by a genotype comprising alleles at L unlinked loci, where a mutation event at any locus always introduces a novel allele. In both the real and simulated MLST data sets considered in the Results section the number of loci L is invariably equal to seven. Recombination between randomly chosen genotypes occurs at any locus according to a Poisson process with the rate defined as $\rho = rN\tau$, where r is the rate per locus in relation to the mutation rate. This is the definition typically employed in bacterial population genetics and it quantifies the recombination rate as the expected number of recombination events per mutation event to reflect the average level of clonality in the population. Given the design of the MLST typing schemes, we assume that each recombination event involves only a single locus. This is motivated by the genomic distances between the chosen loci, which are large enough to have a negligible probability for a single recombination event affecting simultaneously multiple MLST loci.

Microepidemics are modeled as doubly stochastic events, with the frequency of new microepidemics per generation following a Poisson distribution with mean $\omega N\tau$. The size of each microepidemic has a Poisson distribution with mean γ. Each microepidemic is generated independently similar to the assumptions in Fraser et al. such that first a single individual is randomly chosen, after which its genotype is propagated to Y randomly chosen other individuals such that Y has Poisson distribution with mean γ. When the population is stratified, the microepidemic rates of the subpopulations are denoted by ω_1, γ_1 and ω_2, γ_2, respectively. Migration between subpopulations is a Poisson process with the rates $\tau N_1 m_{12}$, $\tau N_2 m_{21}$ per generation, where the first subindex of the parameters m_{12}, m_{21} defines the source and the second subindex the target subpopulation. In migration events genotypes of a Poisson distributed number of randomly chosen individuals from the source population replace the genotypes of randomly chosen individuals in the target population.

In our simulations the model events were generated in the following order: reproduction, mutation, recombination, microepidemics and migration at each generation. The population was simulated until allelic diversity reached equilibrium and in all the reported results each subpopulation size was $N = 2000$, unless otherwise indicated. The allelic mismatch distribution corresponding to state of a population is defined by considering the fractions of all pairs of genotypes which have exactly $l = 0, \ldots, L$ distinct alleles. Medians and 95% confidence intervals for the allelic mismatch distributions were obtained by recording the population state every 100th generation after initial 500 generations until 20000 generations, and using these values to calculate the corresponding quantiles of the mismatch probabilities.

Data and population summaries

MLST isolate data were accessed (September 15, 2014) from the following databases: http://efaecalis.mlst.net/ (*E. faecalis*), http://efaecium.mlst.net/ (*E. faecium*), http://saureus.mlst.net/ (*S. aureus*), and (May 10, 2015) from: http://sepidermidis.mlst.net/ (*S. epidermidis*). These open MLST databases comprise sequences for seven unlinked housekeeping loci, which are under strong purifying selection. Hence, all mutations observed in the corresponding bacterial populations can be assumed to be selectively neutral. The length of these loci varies between approximately 350-550 nucleotides and the total number of third codon positions was used to scale the mutation rate in the analyses. Each observed human isolate was classified as either a commensal or hospital strain based on the available metadata. Isolates with an unknown origin were excluded from the analyses. MLST databases curate and accept submissions of novel data from epidemiological studies, which may be collected locally or globally. Hence, in terms of sampling the global database can be biased by a smaller number of large studies inflating the frequencies of certain genotypes. To mitigate against such biases, we used frequency independent summaries in the model fitting (see below for details).

For a compact visualization of the population data, eBURST networks were produced using default settings (Feil et al., 2004). Turner et al. demonstrated that eBURST provides a robust recapitulation of the genetic relatedness of strains in a bacterial population based on the MLST resolution (Turner et al., 2007). More generally, to quantify topological properties of the network of unique genotypes obtained from a population state and to do inference about model parameters we calculated genotype degree distributions and distributions of geodesic distances between pairs of genotypes, which are standard measures of network topology (Goh et al., 2002). The network of unique genotypes has the advantage over allelic mismatch distribution that each genotype is only considered once, which makes it more robust to oversampling of identical genotypes often occurring in pathogen studies.

The topological summaries were calculated as follows. First, the set of unique genotypes is identified for an observed state of a population (either real or simulated). Then, an adjacency matrix is defined for the unique genotypes, such that a pair of genotypes are adjacent if they differ at maximally one locus. This matrix defines a network for the genotypes from which topological summaries can be calculated. The degree distribution corresponds to the empirical distribution of the number of neighbors a network node has. The geodesic distance between two nodes is the shortest path from a genotype to another genotype in the network. The distribution of geodesic distances is then the empirical distribution calculated over all pairs of genotypes. Here we used the convention of considering only connected network components when calculating the distances and ignoring the genotypes that are isolated from all the remaining genotypes.

Inference

Since the microepidemic models have intractable likelihoods, we used a likelihood-free inference procedure akin to Approximate Bayesian Computation (ABC) inference to fit models to the MLST data with the two topological summary statistics defined previously. The allelic mismatch distributions were not considered in the inference due to concerns about robustness with respect to the mixed sampling strategies involved in the global MLST databases as explained earlier. Depending on the parameter values, simulation of the population until convergence can be very time-consuming and additionally one needs to assess the stochasticity of the summaries for given parameter values. Hence we used the recent Bayesian optimization based likelihood-free inference method (BOLFI) which is several orders of magnitude faster than standard Monte Carlo sampling based ABC inference approaches (Gutmann, 2015). To obtain point estimates for the parameters we used in total 1,000 acquisition points in the parameter space and for each point 100 forward simulations were performed to calculate the expected values of the two summary statistics. The objective function used in BOLFI was the sum of absolute relative errors to the means of the corresponding summaries calculated from each population in the MLST data sets. To investigate the model fit to the data we produced 100 predictive simulations for each data set using values of the point estimates and compared the empirical means with the predictive distribution.

When fitting the models to the data we used the following existing estimates of the recombination rate (*r*/*m*) from the literature: *S. aureus* – 0.43, *S. epidermidis* – 2.5, *E. faecalis* – 0.60, *E. faecium* – 0.70 (de Been et al., 2013a; Everitt et al., 2014; Miragaia et al., 2007; Vos and Didelot, 2009). Since the existing mutation rate estimates are calculated per site and year, their direct use is not feasible in our simulation model as the exact relation between the generation time in the model and in calendar time is unknown. Therefore, we defined the mutation rate in the model as an unknown parameter scaled by the number of third codon positions present in the concatenated MLST gene sequences and a baseline rate equal to $1.5 \cdot 10^{-6}$. The unscaled mutation rate parameter was restricted to the range $10^{-1} – 10^{-3}$, since larger values resulted in extremely slow simulations and smaller values produced an insufficient amount of variation in preliminary runs. In model fitting we used a simplified model with the migration rates set equal to zero, which was equivalent to fitting a three parameter model separately to the commensal and hospital samples for all four species. The two microepidemic rate parameters were both assigned ranges between 10^{-4} and 50.

Results

Simulated populations

We extended the microepidemic infinite alleles model with mutation and recombination rates previously proposed by Fraser et al. (Fraser et al., 2005) to incorporate population stratification, whereby genotypes are free to move between subpopulations at a defined rate. In addition, rather than using a single microepidemic parameter to describe localized transmission (Fraser et al., 2005), we introduced two parameters modulating the distributions of both the frequency and sizes of the transmission clusters in stochastic fashion. Our microepidemic infinite alleles migration model (MIAMI) can thereby encompass a wide variety of evolutionary and ecological parameter space. Since the resulting patterns of genetic variation reflect a complex function of several factors, we consider first a model without population stratification to delineate the influence of each of the model components.

The frequency distribution of the number of allelic mismatches between pairs of genotypes is a classical approach to describe the distribution of genetic variation within a population, see Introduction. Depending on the interplay of several factors, a population may either have a peaked or flat equilibrium distribution over the space of summary statistics, such as the allelic mismatch distribution (Fig. 1). For lower mutation rates, high recombination rate (*r*/*m*) will lead to bell-shaped mismatch distributions, since recombination acts as a cohesive force keeping genetic variation together as a cloud in the space of possible genotypes (Fraser et al., 2007). The mismatch distribution becomes less sensitive to changes in the recombination rate and the equilibrium distribution becomes more peaked when the mutation rate increases (Fig. 1).

Fig. 2 shows the impact of heightened localized transmission (microepidemics) on genetic relatedness visualized using eBURST (Feil et al., 2004; Francisco et al., 2009) and the allele mismatch distribution. The rate of mutation and homologous recombination varies among bacterial pathogens and this can have a marked effect on the population structure. To model the interplay of these two important factors at different levels, four evolutionary scenarios were considered: low mutation and recombination rate (A), mutation dominates (B), recombination dominates (C), both mutation and recombination effects are sizeable (D). If mutation dominates over recombination (Fig. 2,B), microepidemics do not lead to as pronounced changes in the relatedness pattern as in the situation where both mutation and recombination rates are low (Fig. 2,A). Interconnected clusters do emerge under a high rate of recombination, often spanning across large parts of the entire population (Fig. 2,C). The variability of the mismatch distribution at the equilibrium becomes elevated under all regimes of baseline parameter values when microepidemics occur at a frequent rate, as illustrated by the broader confidence intervals (Fig. 2,A-D). Both the frequency and size distribution of the individual microepidemics influence how much probability mass is shifted towards identical genotypes, but the change is also influenced by mutation and recombination rate parameters (Supplementary Fig. 1).

The effect of migration rate on the allelic mismatch distribution within a subpopulation is a complicated function of mutation, recombination and microepidemic rates in a structured population, even if there are only two subpopulations (Fig. 3). We studied the combinations in which a subpopulation undergoes microepidemic expansions at a moderate rate and is coupled with another subpopulation where the rate varies from zero to twice that of the first subpopulation. An increase of the migration rate between the two subpopulations by an order of magnitude leads either to a substantial decrease of the genotypic diversity (Supplementary Fig. 2, i), an increase in the genotypic diversity (Supplementary Fig. 2, a), or to no change at all (Supplementary Fig. 2, e), depending on whether the subpopulation considered as a source experiences more, less, or an equal amount of the microepidemics, compared with the target subpopulation. The effect of migration remains equally complex for the between-subpopulations allelic mismatch distribution, which is insensitive to a change in the migration rate by an order of magnitude for many combinations of subpopulation dynamics (Supplementary Fig. 3). Population stratification combined with asymmetric migration rates can produce patterns of relatedness which are otherwise unlikely under the neutral model (Supplementary Fig. 4). For example, in all our simulations a characteristic U-shaped allelic mismatch distribution only arose when the migration rate was highly asymmetric and one subpopulation experienced considerable microepidemics while the other one had none (Supplementary Figures 5,6).

To obtain some descriptive analytical insight to the joint effect of microepidemic and migration rates on genotypic diversity, we considered how the baseline equilibrium probability of identical genotypes is affected by introducing at the limit a change to the population based on either mechanism. Fraser et al. derived the equilibrium probability of identical genotypes at L unlinked loci, under the assumption of no microepidemics (Fraser et al., 2005), which equals $p_0^L = \frac{1+L\rho p_0^{L-1} p_0^1}{1+L\theta+L\rho}$. Here $\theta = 2\mu N$, where μ is the per locus mutation rate and N is the population size. Furthermore, the recombination rate is defined as $\rho = 2rN$, where r is the rate per locus in relation to the mutation rate. Since this extension of the classical equilibrium result by Kimura to allow for recombination is based on the assumption that in any generation only a single event occurs, Fraser et al. handled the effect of microepidemics on a population at equilibrium implicitly by introducing a probabilistic mixture where a single parameter represents the increase in the probability p_0^L caused by microepidemics. To gain some descriptive insight, we quantify the change in the probability of identical strains by evaluating the expectation of the effect of microepidemic and migration events when allowed only at the equilibrium of a simpler population model with only mutation and recombination events.

Consider first the effect of stochastic microepidemics occurring in a single generation. The expected number of identical genotype pairs arising from them equals $(\gamma + 1)^2 N\omega$, where ω is the scaled rate at which microepidemics occur per generation and γ is the expected size of each microepidemic (Methods). After this single generation the additional fraction of identical strains introduced to the population is $\frac{(\gamma+1)^2 N\omega}{\binom{N}{2}}$, which is an increasing function of both the expected size and rate of microepidemics. Note that this value does not in general equal the probability p_0^L, because the latter depends also on the number of identical genotype pairs already present in the population before introduction of the microepidemic.

Next, consider two subpopulations of sizes N_1, N_2, which at equilibrium become connected with migration rates $N_1 m_{12}$, $N_2 m_{21}$, respectively, in addition to the effect of introducing microepidemics (Methods). Each subpopulation is assumed to have its own set of parameters $\gamma_1^2 N_1 \omega_1, \gamma_2^2 N_2 \omega_2$ governing the extent of microepidemics. Assume now that the subpopulations are of equal size $N_1 = N_2$. Then, the expected contribution to the fraction of identical strains in subpopulation 1 by an increase in the migration rate m_{21} depends generally on whether $\gamma_1^2 N_1 \omega_1 > \gamma_2^2 N_2 \omega_2$ or $\gamma_1^2 N_1 \omega_1 < \gamma_2^2 N_2 \omega_2$, since larger and more frequent microepidemics in subpopulation 2 will increase the probability that the genotypes migrating to subpopulation 1 are identical to each other. Conversely, increased migration from subpopulation 2 will have expected effect of decreasing the probability when the extent of microepidemics in subpopulation 2 is smaller than in subpopulation 1. A difference in the sizes of the subpopulations can further amplify these effects since the rates of events are relative to them. The varying magnitudes of these effects are illustrated by simulation in the Supplementary Figures 2,3,4. However, it should be noted that the stated relationships are dependent on the values of the rate parameters and need not hold throughout the parameter range.

MLST data

Global surveillance data based on MLST typing for several common nosocomial bacterial pathogens (*S. aureus, S. epidermidis, E. faecalis, E. faecium*) generally match well with the expected shape of the allelic mismatch distribution for the considered archetypical population types (Fig. 4). eBURST diagrams provide additional insight into the structure of these populations (Fig. 5). *S. aureus* is known to have a low recombination rate (Everitt et al., 2014) and its population structure is mainly shaped by a combination of mutation rate and intensive clonal expansion of distinct genotypes (Fig. 5, C). Conversely, its sister species *S. epidermidis* displays the bell-shaped mismatch distribution typical for organisms with high recombination rate (Meric et al., 2015) (Fig. 4, D) and a large connected network of related genotypes (Fig. 5, D).

Contrasting the population structures of *E. faecium* and *E. faecalis* reveals marked differences, where *E. faecium* forms large networks of related genotypes characteristic of highly recombinogenic bacteria (Fig. 5, B) (Turner et al., 2007), despite a relatively low estimated recombination rate. *E. faecalis* shows only limited clustering of genotypes (Fig. 5, A) and a mismatch distribution typical for a population dominated by mutation, with a slight increase of identical genotype pairs due to localized hospital transmission (Fig. 4, A).

As shown previously, our analyses of simulated populations under a wide range of microepidemic and migration rates revealed that the allelic mismatch distribution can be insensitive to changes in parameter values over 1-2 orders of magnitude. Hence, we concluded that inferring the model parameters using the allelic mismatch distribution may not be robust enough. Instead, a simplified model with unknown mutation and microepidemic rates was fitted separately to the commensal and hospital populations for each species using the topological summary statistics of the genotype networks (Table 1).

Overall, the simplified neutral model without explicit migration rates fits relatively well to the global MLST data. In particular, for *S. aureus* the observed summary statistics are near the averages of the predictive distribution for both commensal and hospital data, indicating that the characteristic patterns of several disconnected and large clonal complexes can frequently arise under neutrality when the mutation rate is sufficiently low and the microepidemic rates are high. Also for *S. epidermidis* the model fit appears reasonable, albeit the observed mean geodesic distance is higher in the commensal than predicted by the neutral model. The *E. faecalis* population shows somewhat worse fit than *S. epidermidis* and the estimated parameters are all at the end of their ranges in the hospital population, while two estimated parameters are at the end of their ranges in the commensal population. Inspection of the parameter values and the diversity of the observed data suggests that the model attempts to compensate the above bounded mutation rate by strongly increasing the rate of microepidemics while keeping their sizes minimal. The observed *E. faecalis* population has a high allelic diversity which could be explained by a higher mutation rate or effective circulation of alleles from multiple ecological sources by recombination. The only species which shows a more marked deviation from the neutral model is *E. faecium*. In its hospital population the model cannot predict the observed large average geodesic distance resulting from highly interconnected genotypes (Fig 5, D), despite of having a reasonable agreement with respect to the mean degree of the genotypes. The observed values of the two summary statistics are several times larger in the hospital population compared with the commensal population. The model over-predicts their values for the commensal population to an extent, which may be caused by the two parameters that are at their boundaries similar to the *E. faecalis* population.

Table 1. Population characteristics of genotype relatedness for real and simulated data under estimated parameter values. *N* denotes the MLST sample size for each analyzed population.

	S. aureus	*S. epidermidis*	*E. faecalis*	*E. faecium*
N commensal	555	120	225	126
N hospital	543	264	1003	1534
Mean degree commensal	1.57	1.28	0.85	0.81
Mean degree hospital	2.66	1.62	1.04	4.25
Mean geodesic distance commensal	2.02	3.32	1.58	1.46
Mean geodesic distance hospital	2.04	2.68	1.62	4.08
Parameter estimates commensal	$\theta = 0.054$ $\omega = 10^{-4}$ $\gamma = 17.23$	$\theta = 0.165$ $\omega = 11.19$ $\gamma = 38.44$	$\theta = 0.165$ $\omega = 26.80$ $\gamma = 10^{-4}$	$\theta = 0.165$ $\omega = 44.62$ $\gamma = 10^{-4}$
Parameter estimates hospital	$\theta = 0.006$ $\omega = 34.82$ $\gamma = 45.61$	$\theta = 0.054$ $\omega = 10^{-4}$ $\gamma = 13.97$	$\theta = 0.165$ $\omega = 50$ $\gamma = 10^{-4}$	$\theta = 0.012$ $\omega = 11.67$ $\gamma = 46.45$
Commensal:				
Predictive mean degree	1.517	1.368	1.046	1.008
0.1 quantile	1.429	1.220	0.991	0.937
0.9 quantile	1.604	1.605	1.084	1.062
Hospital:				
Predictive mean degree	2.548	1.685	1.023	3.673
0.1 quantile	2.000	1.566	0.977	3.055
0.9 quantile	3.384	1.790	1.069	4.390
Commensal:				
Predictive mean geodesic distance	2.118	2.458	1.706	1.698
0.1 quantile	1.935	2.035	1.619	1.600
0.9 quantile	2.310	3.023	1.779	1.787
Hospital:				
Predictive mean geodesic distance	2.064	2.309	1.728	3.094
0.1 quantile	1.604	2.078	1.649	2.467
0.9 quantile	2.363	2.681	1.800	3.777

Discussion

Previously described neutral models specified by mutation and recombination rate in combination with microepidemics show a close fit to observed local genotype survey data for several commensal and pathogenic bacteria. This holds true for both short-term population evolution dominated by the local dynamics of microepidemics (Fraser et al., 2005; Hanage et al., 2006) and for longer time scales where recombination acts as a cohesive force keeping populations together(Fraser et al., 2007). However, there is limited knowledge about how varying levels of isolation in host organisms, such as human and different animal species (Fraser et al., 2009), might influence the evolutionary dynamics and lead to structured populations in bacteria. Here we considered a neutral model incorporating microepidemics and migration, which mimics a situation where ecological factors limit transmission between subpopulations. Studying the model behavior with extensive simulations, we concluded that allelic mismatch distribution is an insensitive summary statistic under more complex population evolution scenarios where migration and microepidemic rates may vary substantially without notable changes in the population distribution of the summary statistic.

The observed differences between *E. faecium* and *E. faecalis*, which both colonize the gastrointestinal tract, are particularly interesting since mutation and recombination rates have been estimated to be similar for the two species based on both MLST and whole-genome data(de Been et al., 2013b; Vos and Didelot, 2009). Moreover, they are responsible for roughly equal frequencies of nosocomial infections worldwide (Tedim et al., 2015; Willems et al., 2012). *E. faecalis* population structure bears the hallmarks of either a high rate of mutation or drift (or both). *E. faecalis* is known to colonize the vast majority of normal hosts within a population (Tedim et al., 2015), and therefore can be considered as part of the physiological commensal microbiota of humans and many other animals. Certainly, its population structure could be reflective of the evolutionary dynamics of a generalist organism which regularly experiences a high level of drift and gene flow between different host species.

On the basis of the predictions made by our model, *E. faecium* would need to have substantially higher recombination rate than *E. faecalis* to lead to the observed pattern of genotype relatedness under neutrality (Table 1). Since there is evidence of the recombination rate not being substantially higher in *E. faecium* compared with *E. faecalis*, the only possibility for the large genotype networks to arise under our neutral model would be unobserved population stratification. If unobserved sources experiencing very large clonal expansions contributed continuously to the hospital subpopulation of *E. faecium*, the expected allelic mismatch distribution would bear the characteristics of a subpopulation with high recombination rate (Supplementary Fig. 3, i). It is known that intensive farming and animal production practices provide opportunities for rapid clonal expansion of bacterial strains colonizing the animal hosts. Given the known connection between strains from domesticated animals and the hospital associated *E. faecium* (Lebreton et al., 2013; Willems et al., 2012), it is plausible that these clonal expansions could manifest themselves as connected networks in the human hospital subpopulation. However, the extensively connected network of *E. faecium* genotypes would still remain unlikely unless the rate of recombination was substantial. An alternative explanation for the extensive genotype relatedness is a marked deviation from neutrality, such that the connected strains represent either a subpopulation adapted to the hospital environment, consistent with previous studies (Lebreton et al., 2013; Willems et al., 2012), or an adaptation to different host subpopulations (Faith et al., 2015). Further dense sampling will be required to characterize mechanistically the role of hospital adaption for creating the observed relatedness patterns of *E. faecium* strains.

S. aureus and *S. epidermidis* frequently colonize the skin, soft tissue and the nares of human hosts, while also being ubiquitous in a range of animals. However, the overall population density and the proportion of human or animal hosts colonized by *S. epidermidis* largely exceed that of *S. aureus*, so that *S. epidermidis*, but not *S. aureus*, can be considered of a physiological commensal, part of the normal microbiota. Despite of this, *S. aureus* population showed the clearly best fit to the neutral model for both commensal and hospital samples. The human *S. aureus* population is characterized by several genetically distinct clonal complexes, each sharing a single ancestral genotype. Such a population can arise under the neutral mutation/drift driven evolutionary trajectory combined with a high rate of localized transmission, as evidenced by the estimated model. In this scenario clonal complexes appear and proliferate for a time, to be replaced by others arising through genetic drift at the operational timescale of decades or longer. This has been previously described as an 'epidemic clonal' structure (Smith et al., 2000).

Both the commensal and hospital subpopulations of *S. epidermidis* display a pattern of genetic relatedness typical of a population where recombination is the dominant force generating population structure. An exception to this can be seen in the higher fraction of maximally distinct commensal genotypes, which could plausibly arise when novel strains infrequently migrate to the human commensal population from several non-overlapping zoonotic sources (Meric et al., 2015).

The complexities of within- and between-subpopulation strain dependence, and the extent of localized transmission and migration across ecological patch boundaries makes formal statistical inference about microepidemics and migration rates difficult. Given the sampling limitations of the MLST data and the computational challenges, we abstained from fitting the full model with non-zero migration rates to the considered pathogen species. A particular challenge is that, when a population evolves within a drift dominated model, it is unlikely that reliable estimates of the parameters driving the population dynamics can be obtained without dense longitudinal sampling, since cross-sectional samples from the population structure can vary substantially. Similarly, as the consequences of migration events are dependent on other stochastically varying factors across the subpopulations, high migration rates may lead to a pattern of relatedness indistinguishable from those generated by low rates. It is possible that these issues could be resolved using various coalescent-based models, including the Λ-coalescent and its generalizations (Beerli and Felsenstein, 1999; Beerli and Felsenstein, 2001; Choi and Hey, 2011; Hey and Machado, 2003; Hey and Nielsen, 2004; Pitman, 1999; Sagitov, 1999), which motivates further research on their adaptation to the study of large-scale bacterial pathogen populations.

It is evident that a limited number of neutrally evolving core genes, such as those typically used in the MLST typing schemes, limits the scope of models that can be fitted to genetic surveillance data. However, our results imply that some evolutionary scenarios would remain unidentifiable even if housekeeping loci were considered at the whole-genome scale, in particular if the data are mainly cross-sectional even if densely covering the host population. Hence, one of our main conclusions is that the optimal data for studying dynamics in this fashion are densely sampled longitudinal surveillance data covering evolutionary events at whole-genome level (Croucher et al., 2013). This highlights the importance of easy access online repositories of genomic variation as an extension of the currently existing MLST databases and that sample metadata should be an equally important focus of the data sharing principles. Using such a strategy in the near future may enable important model-based predictions about the dynamics of existing and emerging pathogens that pose a considerable global challenge for human and animal health.

Acknowledgments

J.C., E.N., M.G. and M.S. were funded by the grant 251170 from Academy of Finland.

Author contributions

J.C., E.N., M.G. developed and implemented the model, P.M. and M.S. provided additional expertise for the model development and analyses, J.C, G.M., S.K.S, T.C., F.B., W.V.S., R.W., E.F., W.P.H. provided data, biological expertise and interpretation, J.C., E.N., E.F. and W.P.H. wrote the manuscript. All authors approved the final manuscript.

References

Bamshad, M. J., Mummidi, S., Gonzalez, E., Ahuja, S. S., Dunn, D. M., Watkins, W. S., Wooding, S., Stone, A. C., Jorde, L. B., Weiss, R. B., Ahuja, S. K., 2002. A strong signature of balancing selection in the 5' cis-regulatory region of CCR5. Proc Natl Acad Sci U S A 99, 10539-44, doi:10.1073/pnas.162046399.

Beerli, P., Felsenstein, J., 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics 152, 763-73.

Beerli, P., Felsenstein, J., 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proc Natl Acad Sci U S A 98, 4563-8, doi:10.1073/pnas.081068098.

Choi, S. C., Hey, J., 2011. Joint inference of population assignment and demographic history. Genetics 189, 561-77, doi:10.1534/genetics.111.129205.

Croucher, N. J., Finkelstein, J. A., Pelton, S. I., Mitchell, P. K., Lee, G. M., Parkhill, J., Bentley, S. D., Hanage, W. P., Lipsitch, M., 2013. Population genomics of post-vaccine changes in pneumococcal epidemiology. Nature Genetics 45, 656-+, doi:Doi 10.1038/Ng.2625.

de Been, M., van Schaik, W., Cheng, L., Corander, J., Willems, R. J., 2013. Recent recombination events in the core genome are associated with adaptive evolution in Enterococcus faecium. Genome Biol Evol 5, 1524-35, doi:10.1093/gbe/evt111.

Everitt, R. G., Didelot, X., Batty, E. M., Miller, R. R., Knox, K., Young, B. C., Bowden, R., Auton, A., Votintseva, A., Larner-Svensson, H., Charlesworth, J., Golubchik, T., Ip, C. L., Godwin, H., Fung, R., Peto, T. E., Walker, A. S., Crook, D. W., Wilson, D. J., 2014. Mobile elements drive recombination hotspots in the core genome of Staphylococcus aureus. Nat Commun 5, 3956, doi:10.1038/ncomms4956.

Ewens, W. J., 2004. Mathematical population genetics. Springer, New York.

Faith, J. J., Colombel, J. F., Gordon, J. I., 2015. Identifying strains that contribute to complex diseases through the study of microbial inheritance. Proceedings of the National Academy of Sciences of the United States of America 112, 633-640, doi:DOI 10.1073/pnas.1418781112.

Feil, E. J., Li, B. C., Aanensen, D. M., Hanage, W. P., Spratt, B. G., 2004. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. J Bacteriol 186, 1518-30.

Francisco, A. P., Bugalho, M., Ramirez, M., Carrico, J. A., 2009. Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. BMC Bioinformatics 10, 152, doi:10.1186/1471-2105-10-152.

Fraser, C., Hanage, W. P., Spratt, B. G., 2005. Neutral microepidemic evolution of bacterial pathogens. Proc Natl Acad Sci U S A 102, 1968-73, doi:10.1073/pnas.0406993102.

Fraser, C., Hanage, W. P., Spratt, B. G., 2007. Recombination and the nature of bacterial speciation. Science 315, 476-80, doi:10.1126/science.1127573.

Fraser, C., Alm, E. J., Polz, M. F., Spratt, B. G., Hanage, W. P., 2009. The bacterial species challenge: making sense of genetic and ecological diversity. Science 323, 741-6, doi:10.1126/science.1159388.

Goh, K. I., Oh, E., Jeong, H., Kahng, B., Kim, D., 2002. Classification of scale-free networks. Proc Natl Acad Sci U S A 99, 12583-8, doi:10.1073/pnas.202301299.

Gutmann, M. U., Corander, J., 2015. Bayesian Optimization for Likelihood-Free Inference of Simulator-Based Statistical Models. Journal of Machine Learning Research In press, arXiv:1501.03291v3.

Hanage, W. P., Fraser, C., Spratt, B. G., 2006. The impact of homologous recombination on the generation of diversity in bacteria. J Theor Biol 239, 210-9, doi:10.1016/j.jtbi.2005.08.035.

Harpending, H. C., 1994. Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. Hum Biol 66, 591-600.

Hartl, D. L., Clark, A. G., 2007. Principles of population genetics. Sinauer Associates, Sunderland, Mass.

Hey, J., Machado, C. A., 2003. The study of structured populations--new hope for a difficult and divided science. Nat Rev Genet 4, 535-43, doi:10.1038/nrg1112.

Hey, J., Nielsen, R., 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D. persimilis. Genetics 167, 747-60, doi:10.1534/genetics.103.024182.

Hudson, R. R., 1987. Estimating the recombination parameter of a finite population model without selection. Genet Res 50, 245-50.

Lebreton, F., van Schaik, W., McGuire, A. M., Godfrey, P., Griggs, A., Mazumdar, V., Corander, J., Cheng, L., Saif, S., Young, S., Zeng, Q. D., Wortman, J., Birren, B., Willems, R. J. L., Earl, A. M., Gilmore, M. S., 2013. Emergence of Epidemic Multidrug-Resistant Enterococcus faecium from Animal and Commensal Strains. MBio 4, doi:ARTN e00534-13

DOI 10.1128/mBio.00534-13.

Meric, G., Miragaia, M., de Been, M., Yahara, K., Pascoe, B., Mageiros, L., Mikhail, J., Harris, L. G., Wilkinson, T. S., Rolo, J., Lamble, S., Bray, J. E., Jolley, K. A., Hanage, W. P., Bowden, R., Maiden, M. C., Mack, D., de Lencastre, H., Feil, E. J., Corander, J., Sheppard, S. K., 2015. Ecological Overlap and Horizontal Gene Transfer in Staphylococcus aureus and Staphylococcus epidermidis. Genome Biol Evol 7, 1313-28, doi:10.1093/gbe/evv066.

Miragaia, M., Thomas, J. C., Couto, I., Enright, M. C., de Lencastre, H., 2007. Inferring a population structure for Staphylococcus epidermidis from multilocus sequence typing data. J Bacteriol 189, 2540-52, doi:10.1128/JB.01484-06.

Mousset, S., Derome, N., Veuille, M., 2004. A test of neutrality and constant population size based on the mismatch distribution. Mol Biol Evol 21, 724-31, doi:10.1093/molbev/msh066.

Pitman, J., 1999. Coalescents with multiple collisions. Annals of Probability 27, 1870-1902.

Plucinski, M. M., Starfield, R., Almeida, R. P., 2011. Inferring social network structure from bacterial sequence data. PLoS One 6, e22685, doi:10.1371/journal.pone.0022685.

Rogers, A. R., Harpending, H., 1992. Population growth makes waves in the distribution of pairwise genetic differences. Mol Biol Evol 9, 552-69.

Rousset, F., 2004. Genetic Structure and Selection in Subdivided Populations. Princeton University Press, Princeton.

Sagitov, S., 1999. The general coalescent with asynchronous mergers of ancestral lines. Journal of Applied Probability 36, 1116-1125.

Smith, J. M., Feil, E. J., Smith, N. H., 2000. Population structure and evolutionary dynamics of pathogenic bacteria. Bioessays 22, 1115-22, doi:10.1002/1521-1878(200012)22:12<1115::AID-BIES9>3.0.CO;2-R.

Tedim, A. P., Ruiz-Garbajosa, P., Corander, J., Rodriguez, C. M., Canton, R., Willems, R. J., Baquero, F., Coque, T. M., 2015. Population biology of intestinal enterococcus isolates from hospitalized and nonhospitalized individuals in different age groups. Appl Environ Microbiol 81, 1820-31, doi:10.1128/AEM.03661-14.

Turner, K. M., Hanage, W. P., Fraser, C., Connor, T. R., Spratt, B. G., 2007. Assessing the reliability of eBURST using simulated populations with known ancestry. BMC Microbiol 7, 30, doi:10.1186/1471-2180-7-30.

Vos, M., Didelot, X., 2009. A comparison of homologous recombination rates in bacteria and archaea. ISME J 3, 199-208, doi:10.1038/ismej.2008.93.

Willems, R. J., Top, J., van Schaik, W., Leavis, H., Bonten, M., Siren, J., Hanage, W. P., Corander, J., 2012. Restricted gene flow among hospital subpopulations of Enterococcus faecium. MBio 3, e00151-12, doi:10.1128/mBio.00151-12.

Wood, S. N., 2010. Statistical inference for noisy nonlinear ecological dynamic systems. Nature 466(7310):1102–1104.

Figure legends

Fig. 1. Allelic mismatch distributions for combinations of mutation and recombination rates in a population with $N = 3000$. Bold line in green shows the mean mismatch probability over 20000 generations, sampled at intervals of 100 generations. The green shaded area shows the 95% confidence interval and the colored lines are examples of mismatch distributions at random time points. Vertical axis in each panel shows the probability mass associated with the points of the curves across the values on the horizontal axis which correspond to the possible allelic distances with seven loci. Distributions are shown as continuous curves for visual clarity.

Fig. 2. eBURST networks and mismatch distributions for a population without (grey) and with (yellow) microepidemics, where $\omega = 27$, $\gamma = 16$. The 95% confidence intervals are shown by shaded areas and are defined as in Fig. 1. Numbers in the networks represent arbitrary genotype labels. The mutation and recombination parameters used are: 0.0011, 1 (A), 0.0088, 1 (B), 0.0011, 8 (C), 0.0088, 8 (D).

Fig. 3. Schematic illustration of the combined effect of microepidemics and migration studied in detail in the supplementary materials. Colors represent genotype clusters. The population on the left is unstratified, in which case increasing rate (ω) and size (γ) of microepidemics lead to decreased genetic variation. In a stratified population with two subpopulations (P_1, P_2) the effect of increasing microepidemics (ω_1, γ_1) on genetic diversity in subpopulation P_1 depends both on the microepidemics in subpopulation P_2 (ω_2, γ_2) and on the migration rate (m_{21}). The case with $m_{21} = 0$ leads to identical decrease of genetic variation as in an unstratified population. The notation "<<" is used to indicate that the parameters on the left side of the double inequality are much smaller than those on the right side.

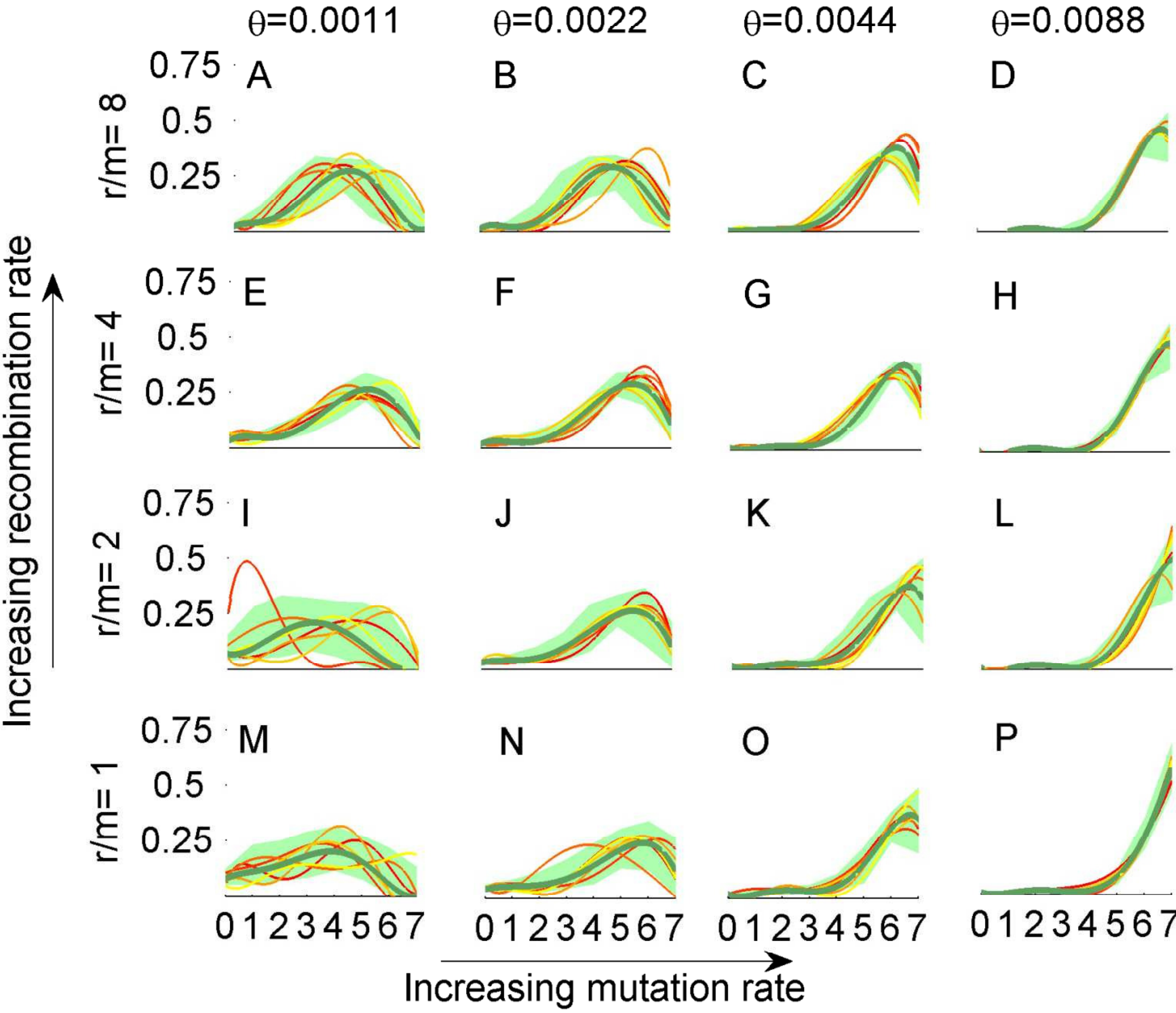
Fig. 4. Allelic mismatch distributions of commensal and hospital subpopulations of four common nosocomial bacterial pathogens. The right-most column shows the between-subpopulation mismatch distributions.

Fig. 5. eBURST networks of the isolates used to calculate the mismatch distribution in Fig. 4; *E. faecalis* (A), *E. faecium* (B), *S. aureus* (C), *S. epidermidis* (D).

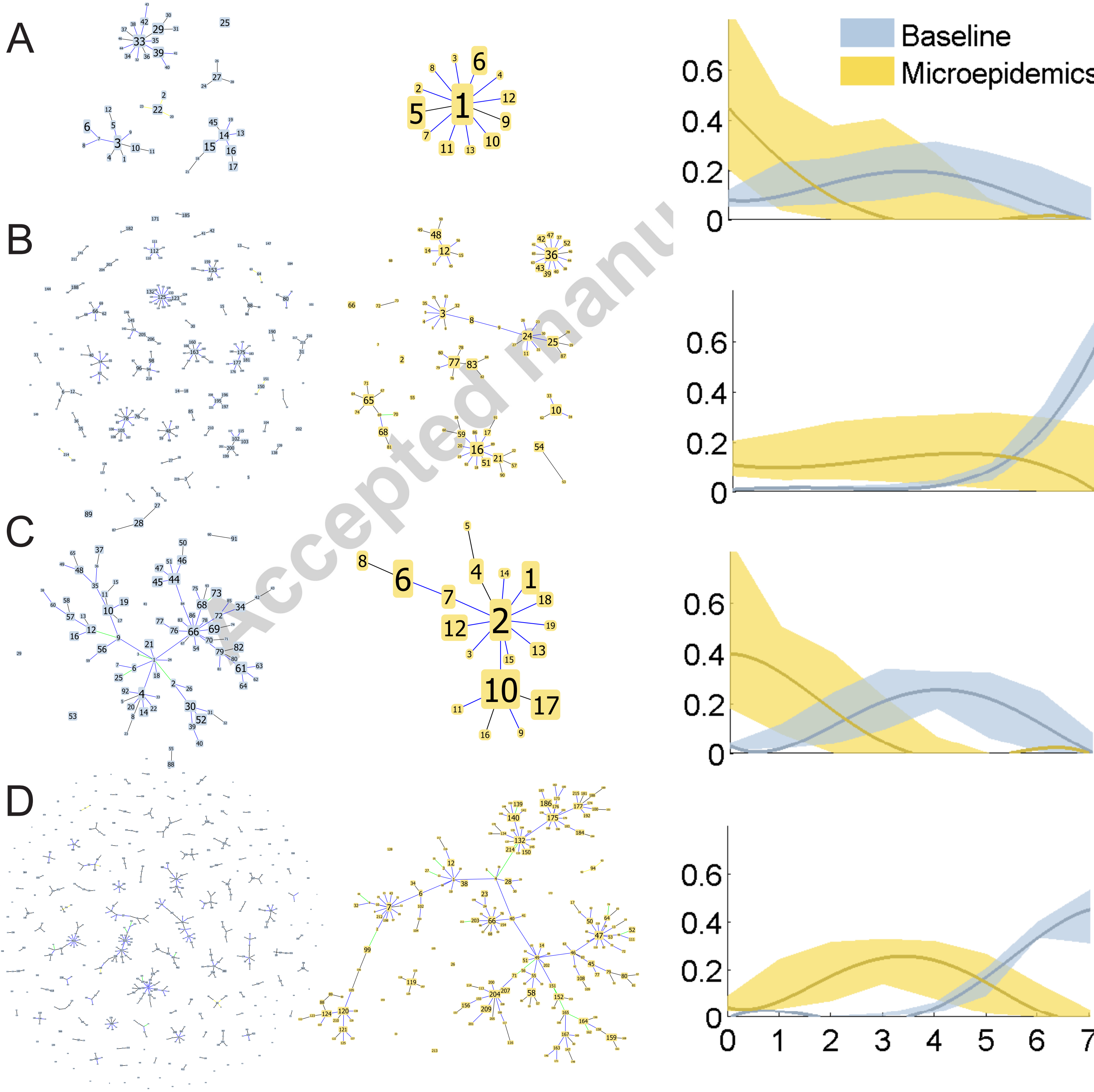
highlights

- Bacterial pathogens pose a considerable global challenge for human and animal health
- The ability to draw robust inferences about evolutionary dynamics is important
- Population stratification and host migration weaken parameter identifiability
- A major hospital pathogen, *Staphylococcus aureus*, fits well with a neutral model
- The hospital pathogen *Enterococcus faecium* shows strong deviation from neutrality

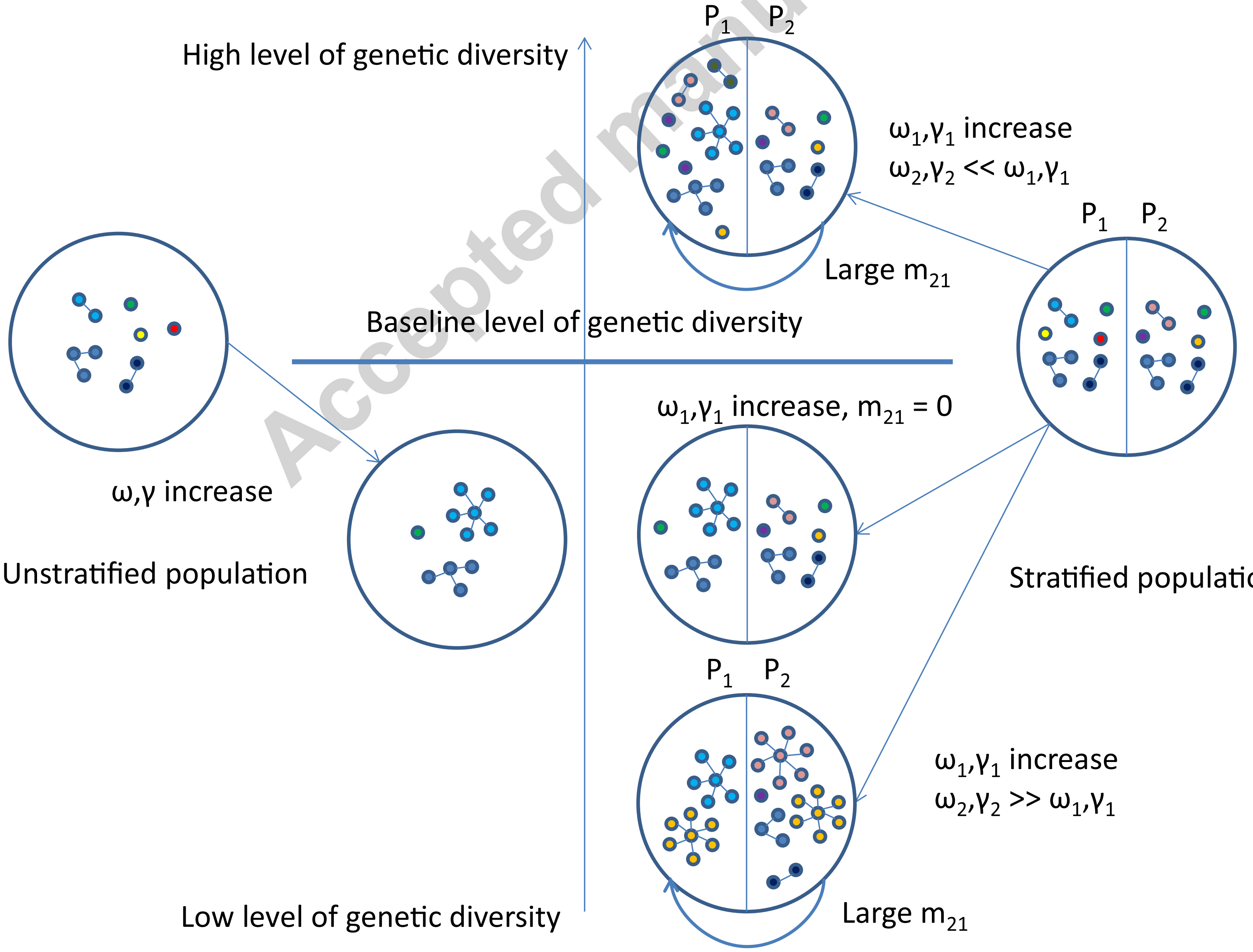
4. Figure


θ=0.0011
θ=0.0022
θ=0.0044
θ=0.0088
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
r/m= 8
r/m= 4
r/m= 2
r/m= 1
0.75
0.5
0.25
0 1 2 3 4 5 6 7
Increasing recombination rate
Increasing mutation rate

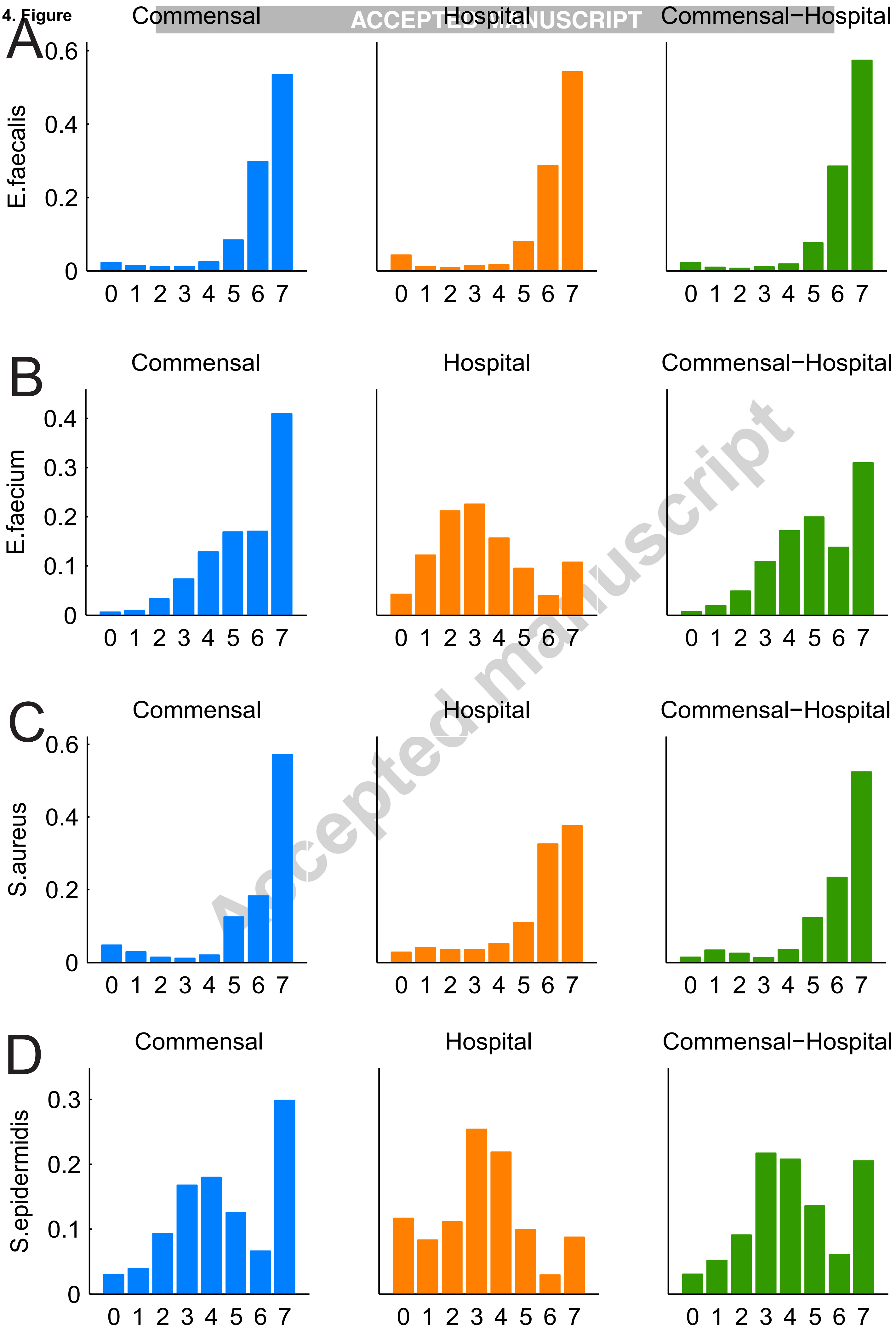
4. Figure



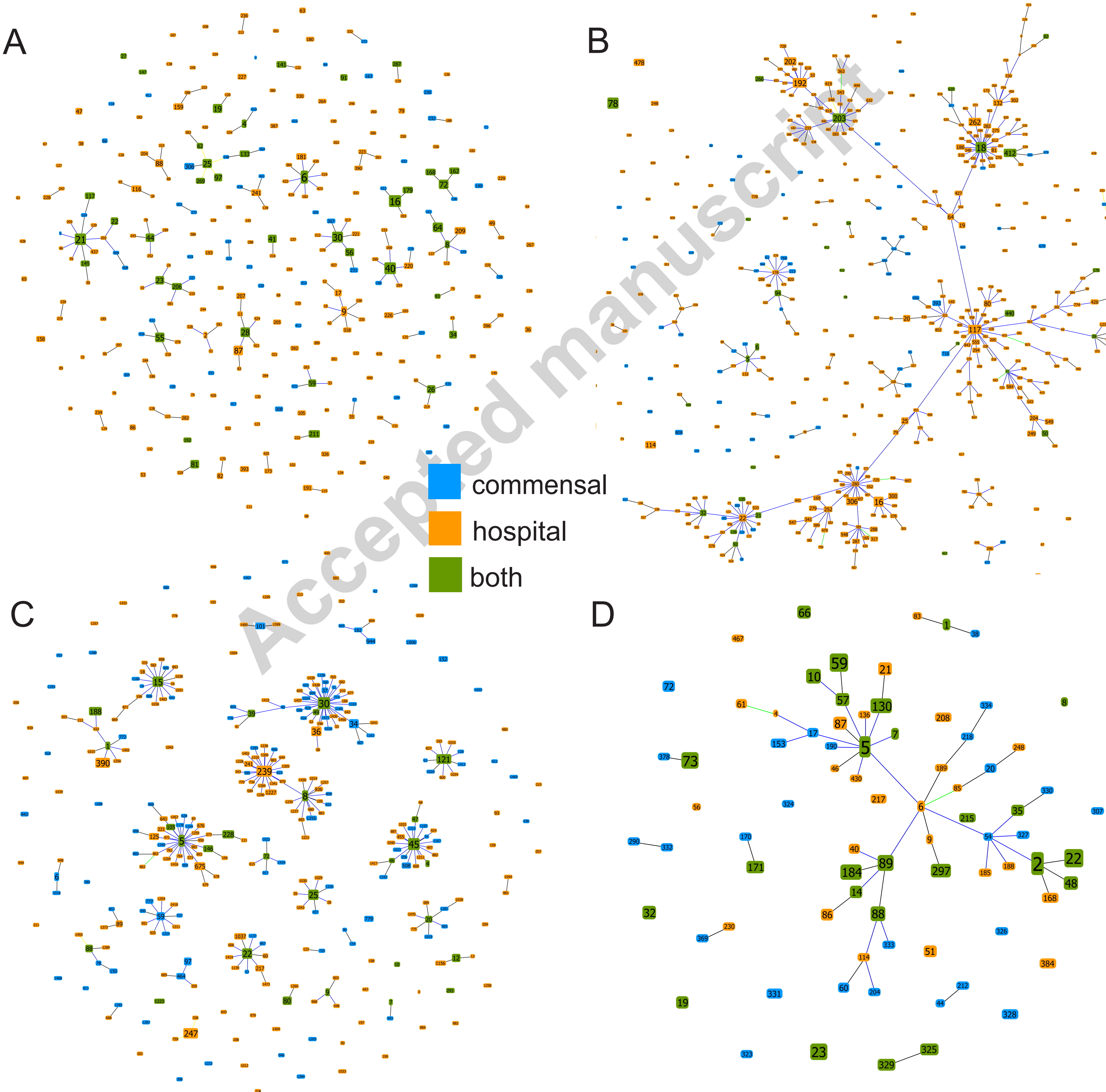
4. Figure


P_1 P_2
High level of genetic diversity
ω_1,γ_1 increase
$\omega_2,\gamma_2 << \omega_1,\gamma_1$
P_1 P_2
Large m_{21}
Baseline level of genetic diversity
ω_1,γ_1 increase, $m_{21} = 0$
ω,γ increase
Unstratified population
Stratified populatio
P_1 P_2
ω_1,γ_1 increase
$\omega_2,\gamma_2 >> \omega_1,\gamma_1$
Low level of genetic diversity
Large m_{21}

4. Figure


A
Commensal
Hospital
Commensal−Hospital
E.faecalis
0.6
0.4
0.2
0
0 1 2 3 4 5 6 7
B
Commensal
Hospital
Commensal−Hospital
E.faecium
0.4
0.3
0.2
0.1
0
0 1 2 3 4 5 6 7
C
Commensal
Hospital
Commensal−Hospital
S.aureus
0.6
0.4
0.2
0
0 1 2 3 4 5 6 7
D
Commensal
Hospital
Commensal−Hospital
S.epidermidis
0.3
0.2
0.1
0
0 1 2 3 4 5 6 7

4. Figure


A
B
C
D
commensal
hospital
both
66
83
1
38
467
59
21
10
72
57
130
61
4
136
8
87
208
334
153
17
7
378
73
190
5
218
248
46
430
189
20
85
56
324
217
6
330
307
215
35
290
332
170
9
54
327
40
22
89
184
297
185
188
2
171
48
14
168
32
88
86
326
230
369
333
51
114
384
60
204
212
331
44
19
328
323
23
325
329