Bayesian phylogenetics of bacterial genomes: inferring the pandemic spread of
Methicillin resistant *Staphylococcus aureus* ST239

Rebecca R. Gray$^{1,2}$, Andrew J. Tatem$^{1,3}$, Judith A. Johnson$^{1,2}$, Alexander V. Alekseyenko$^4$, Oliver G. Pybus$^5$, Marc A. Suchard$^{6,7,8}$, Marco Salemi$^{1,2}$

1. Emerging Pathogens Institute, University of Florida, Gainesville, FL, USA
2. Department of Pathology, Immunology and Laboratory Medicine, University of Florida, Gainesville, FL, USA.
3. Department of Geography, University of Florida, Gainesville, FL, USA
4. Center for Health Informatics and Bioinformatics, New York University School of Medicine.
5. Department of Zoology, University of Oxford, Oxford, UK
6. Department of Biomathematics, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA
7. Department of Biostatistics, UCLA School of Public Health, Los Angeles, CA
8. Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA

Corresponding authors:

Rebecca R. Gray
Dept. of Pathology, Immunology, and Laboratory Medicine
Emerging Pathogens Institute
University of Florida College of Medicine
1376 Mowry Road
Gainesville, FL 32610
Tel. +1 352-273-8164
Fax +1 352-273-8284
rgray@ufl.edu

Marco Salemi
Dept. of Pathology, Immunology, and Laboratory Medicine
Emerging Pathogens Institute

© The Author 2010. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com
Research Article

Key Words: Bayesian inference, phylogeography, phylogenetics, measurably evolving population

Running Head: Pandemic spread of HA-MRSA ST239
ABSTRACT

*Staphylococcus aureus* is a common cause of infections that has undergone rapid global spread over recent decades. Formal phylogeographic methods have not yet been applied to the molecular epidemiology of bacterial pathogens because the limited genetic diversity of datasets based on individual genes usually results in poor phylogenetic resolution. Here, we investigated a whole genome SNP dataset of healthcare-associated Methicillin resistant *S. aureus* sequence type 239 (HA-MRSA ST239) strains, which we analyzed using Markov spatial models that incorporate geographical sampling distributions. The reconstructed timescale indicated a temporal origin of this strain shortly after the introduction of Methicillin, followed by global pandemic spread. The estimate of the temporal origin was robust to the molecular clock, coalescent prior, full/intergenic/synonymous SNP inclusion, and correction for excluded invariant site patterns. Finally, phylogeographic analyses statistically supported the role of human movement in the global dissemination of HA-MRSA ST239, although it was unable to conclusively resolve the location of the root. This study demonstrates that bacterial genomes can indeed contain sufficient evolutionary information to elucidate the temporal and spatial dynamics of transmission. Future applications of this approach to other bacterial strains may provide valuable epidemiological insights that may justify the cost of genome-wide typing.
INTRODUCTION

High-resolution Bayesian phylogenetic methods are powerful tools, capable of elucidating the temporal and spatial dynamics of viral epidemics (Pybus, Rambaut 2009), including those of HIV-1 (i.e. Gray et al. 2009), HCV (i.e. Markov et al. 2009) and influenza (i.e. Rambaut et al. 2008). Because these methods assume a molecular clock, the dataset under study must either represent sequences sampled at different times from a measurably evolving population (MEP; Drummond et al. 2003) or the study must incorporate an independent calibration date or evolutionary rate (Pybus 2006). A population can be considered to be measurably evolving only if a significant number of mutations accumulate between sampling times (Drummond et al. 2002), which can be evaluated using a likelihood ratio test (Rambaut 2000). Until now, these methods have been typically applied to RNA viruses, which accumulate mutations at rates around $10^{-3}$-$10^{-4}$ nucleotide substitutions/site/year, thereby generating sufficient genetic diversity within their relatively short genomes ($\sim 10^5$ nucleotides) over limited periods of time (months, years or decades; Drummond, Pybus, Rambaut 2003; Belshaw et al. 2008). A recent study has evaluated the reliability of using MEP methods when applied to DNA viruses, whose evolutionary rates are thought to be orders of magnitude lower than RNA viruses (Firth et al. 2010). Based on extensive simulations, the authors concluded that, given enough variable sites, accurate estimation of divergence times was possible even when the underlying evolutionary rate was $10^{-7}$ sub/site/year (Firth et al. 2010).

Bacterial populations, on the other hand, have typically not been considered as MEPs, since the limited genetic diversity of individual genes, or of MLST loci, contain insufficient phylogenetic signal (Achtman 2008) to exhibit measurable evolution over the timeframes of typical epidemiological studies. Only three studies to date have used serial-sample coalescent methods to reconstruct evolutionary patterns in bacteria: *Staphylococcus aureus* sequence type (ST)-5 (Lowder et al. 2009) and ST-225 (Nubel et al. 2010) and *Neisseria gonorrhoeae* (Tazi et al. 2010), although none of these studies employed Bayesian phylogeography. The evolutionary
rates estimated for \textit{S. aureus} when the sampling dates of serially-sampled sequences were taken into account were on the order of $10^{-6}$ subs/site/year \cite{harris2010, nubel2010}. In contrast, older estimates of bacterial evolutionary rates, based on external calibration dates, were on the order of $10^{-9}$ subs/site/year \cite{ochman1999}. It is known that using an independent calibration date or evolutionary rate can often lead to erroneous time scales if chosen without diligence \cite{shapiro2006}, as rates of nucleotide change can differ by several orders of magnitude depending on the amount of time separating two sequences \cite{ho2005}. Indeed, the slower rate estimate for bacteria has been acknowledged to suspect by the original author \cite{achtman2008}. On the other hand, is unclear whether heterochronous methods can be applied to bacteria, i.e. whether they represent an MEP. If Bayesian methods are applied inappropriately to serially-sampled sequences then inappropriate conclusions may be drawn \cite{firth2010}.

Recent advances in sequencing technologies means that the investigation of bacterial genetic variation at the whole genome level is now experimentally and economically feasible. This gives rise to the possibility that enough bacterial genetic diversity can be observed over short timeframes to consider such populations as measurably evolving, thereby allowing the estimation of divergence times and the reconstruction of dispersal patterns without the need for an independent temporal calibration. In this study, we investigated a dataset of 4,310 genome-wide single nucleotide polymorphisms (SNPs) of health-care associated Methicillin resistant \textit{S. aureus} (HA-MRSA), present within an alignment of 63 geographically-diverse sequences sampled over a 24-year period \cite{harris2010, see table S1 for strains}. We aimed to determine whether this genome-wide dataset met the requirements for Bayesian coalescent analysis. First, a significant number of nucleotide changes must be observed, thus suggesting strong phylogenetic signal. Second, the dataset must demonstrate low signal for recombination. Third, the amount of time between sampling times must provide enough time for a significant number of mutations to have arisen. These conditions were tested on (i) the full dataset,
comprising 4310 SNPs among 63 taxa, (ii) a subset that included only synonymous changes (1055 SNPs) and (iii) a subset that included only intergenic changes (962 SNPs). Further Bayesian phylogeographic analyses were then performed to investigate the temporal and spatial spread of HA-MRSA ST-239.
METHODS

Datasets. The alignment of 4310 SNPs among 63 isolates was downloaded from the Supplemental files of (Harris et al. 2010). Sub-sampled datasets were constructed as described above. Sequences were named following the convention of (Harris et al. 2010) including the year of sampling (Table S1).

Likelihood mapping. To investigate the phylogenetic signal of each dataset, likelihood mapping was performed using the TREE-PUZZLE program by analyzing 10,000 random quartets (Schmidt et al. 2002). This method proceeds by evaluating, using maximum likelihood, groups of four randomly chosen sequences (quartets). For each quartet, the three possible unrooted tree topologies are weighted. The posterior weights are then plotted using triangular coordinates, such that each corner represents a fully resolved tree topology. Hence, dots localized close to the triangle vertices represent tree-like phylogenetic signal, while those close to the center and on the sides represent star-like (completely unresolved) and network-like (partially unresolved) signal, respectively (Strimmer, von Haeseler 1997).

Recombination. Recombination creates mosaic genomes, which violate the assumption of tree-like evolution. Therefore, a network was inferred for each dataset using SplitsTree (Huson, Bryant 2006). Each dataset was analyzed for the presence of recombinant sequences using the PHI test (Bruen, Philippe, Bryant 2006) with alpha = 0.001 (Salemi, Gray, Goodenow 2008).

Bayesian Phylogenetic Inference. To estimate the phylogeny for each alignment we used the Bayesian framework implemented in BEAST software package version 1.5.4 (Drummond, Rambaut 2007) under the GTR nucleotide substitution model. The molecular clock was calibrated under either a strict molecular clock (which assumes the same evolutionary rates for all branches in the tree) or a relaxed clock (which allows different rates on different branches, drawn from a specified distribution; Drummond et al. 2006b). Both the constant population size coalescent prior or the extended Bayesian skyline plot model (allowing change
in effective population size over time; Heled, Drummond 2010) were tested. The MCMC
analysis was run until evidence of proper mixing was obtained (up to $10^8$ generations, see
below); the chain was sampled every 10000$^{th}$ generation. For analyses assuming the relaxed
clock, three independent runs were combined to achieve proper mixing of the chain. Results
were visualized in Tracer v.1.5, and proper mixing of the MCMC was assessed by calculating
the effective sampling size (ESS) for each parameter (Drummond, Rambaut 2007). All ESS
values were >200. For each dataset, the maximum clade credibility (MCC) tree, which is the
tree with the largest product of posterior clade probabilities, was selected from the posterior tree
distribution (after removal of 50% burnin) using the program TreeAnnotator version 1.5.4
(available as part of the BEAST package). Final trees were annotated with FigTree version 1.3
(http://tree.bio.ed.ac.uk/software/figtree/).

Because only variable sites were used in the analysis, an ascertainment bias correction
(ABC) model was implemented in a subset of analyses. For a nucleotide model with 4 states
and a $N$-taxon tree, there are $4^N$ possible site patterns. Without an ascertainment bias, the sum
of the probabilities of all $4^N$ patterns = 1. If specific types of patterns (such as all invariant sites)
are deliberately excluded, however, then the probabilities of site patterns no longer sum to 1. To
correct for this, the computed probabilities for the observed site patterns can be renormalized,
which is implemented in the trunk source code of BEAST available from

**Evaluation of competing models.** Models were compared by calculating the Bayes
Factor (BF), which is the ratio of the marginal likelihoods (marginal with respect to the prior) of
the two models being compared (Kass, Raftery 1995; Suchard, Weiss, Sincheimer 2001a). We
calculated approximate marginal likelihoods for each coalescent model via importance sampling
(1000 bootstraps) using the harmonic mean of the sampled likelihoods (with the posterior as the
importance distribution). The ratio of the marginal likelihoods between any two models is the
Bayes Factor (BF). Evidence against the null model (i.e. the one with lower marginal likelihood)
is indicated by $2 \ln(BF) > 3$ (positive) and $> 10$ (strong). The calculations were performed with BEAST version 1.5.4 and Tracer v.1.5.

**Bayesian Phylogeographic analysis.** The full dataset was used in this analysis, although the DEN907 strain was excluded because of uncertainly in its location of origin (n=62). Ancestral reconstruction of discrete states was tested in a Bayesian statistical framework implemented in BEAST 1.5.4 (Lemey et al. 2009). A matrix of geographic locations was constructed based on the city of sampling for each sequence. Using the full city dataset resulted in poor mixing of the MCMC chain because the model was over-parameterized; i.e. too little information in the sequence data to accurately estimate the migration matrix among locations. Thus, for sequences that were sampled from multiple cities in the same country (e.g. Hungary, Portugal, Brazil, and Czech Republic), the centroid distance between cities was calculated and used as the sampling location in order to reduce the number of locations. This approach retained the global distribution of the sampling locations while only minimally sacrificing local resolution within a country. This is true even for Brazil, where the two sampled cities (Sao Paulo and Rio de Janeiro) are only ~200 miles apart. A total of 13 sampling locations were used: Capital Federal/Buenos Aires (Argentina), Melbourne (Australia), Brazil, Santiago (Chile), Nanjing (China), Czech Republic, Patras (Greece), Hungary, Portugal, Udon (Thailand), Ankara (Turkey), Montevideo (Uruguay), and New York (USA). Strain TW20 was coded as Udon because of its probable epidemiological link with the Thailand outbreak (Harris et al. 2010). A full model was used in which all 78 possible reversible exchange rates between locations were positive. A mixed model was also tested which uses the Bayesian Stochastic Search Variable Selection (BSSVS) procedure in which exchange rates are allowed to be zero with some probability. Specifically, a truncated Poisson prior that assigns 50% prior probability on the minimal rate configuration, comprising 12 non-zero rates connecting the 13 locations, was assumed (Lemey et al. 2009). Three different priors on the rate matrix were used under both the full and the BSSVS model: a flat prior (equal probability for all rates); a distance informed-prior
which was proportional to the Euclidean distance between cities/centroids; and a migration-
informed prior, proportional to the number of migrants from each sampled county living in any
other sampled county normalized by the total population size.

Files for viewing in Google Earth were created using the MCC treefile, geographic
coordinates for each of the sampled cities/centroids, and a script available at
http://beast.bio.ed.ac.uk. The xml files for all BEAST analyses are available from
http://datadryad.org/.

**Summarizing posterior location uncertainty.** A model of low statistical power makes
poor use of the information in the data, while a successful model exploits this information to
generate posterior distributions that are maximally different from prior beliefs. As in indication of
the amount of information extracted under each mode/prior, the Kullback-Leibler (KL)
divergence (Kullback, Leibler 1951) was calculated for the uncertainty in the ancestral
reconstruction of the root location, using a uniform discrete distribution as reference distribution,
as described by Lemey *et al.* (2009). KL divergence is a non-symmetric measure of the
difference between the prior and the posterior distribution, where larger values signify that a
given Bayesian phylogeographic model extracts more information from the data. KL divergence
values for the root location were compared for analyses using different prior distributions on the
migration matrix (uniform, human migration, inverse distance).

**Geographic information system data.** The great-circle distances (i.e. the shortest
distance between any two points on the surface of a sphere, or orthodromic distance) between
each pair of sample locations were calculated using ArcGIS 9.3. To obtain rates of global
human migration we initially sought data at a sub-national, regional level, in order to capture the
full range of relevant movements. However, such data are non-existent for the vast majority of
countries, and are patchy, unstandardized and variable for the remainder. Foreign-born and
foreign-national population data derived from recent censuses represent the most complete and
comparable datasets for global and regional analyses that most readily accord with actual
population movements (Parsons et al. 2007), and these were used here as a measure of the relative levels of movement between countries.

Data on international bilateral migrant stocks for 226 countries and territories in 2000-2002 were obtained (Parsons et al. 2007). Wherever possible, these data were derived from the latest round of censuses, as these were considered most comparable at the global level. Where unavailable, population registers were drawn upon, and in the cases of missing data, a variety of techniques and tests were employed to create and validate a complete matrix of international bilateral migrant stocks (Parsons et al. 2007). Finally, all data prior to 2000 were scaled to the United Nations mid-year totals of migrant stocks for 2000 (United Nations 2004). For each country or territory, the completed dataset represented the number of foreign-born and foreign-nationality people in residence in 2000-2002, and which country/territory in which they were born or from which they had traveled. It should of course be noted that these data do not capture either very short-term or illegal movements, which themselves can be substantially larger than those in official records. The migrant dataset was rescaled to account for the population size of each country of origin, providing a measure of the strength of migration between countries. For instance, a country containing 100,000 residents born in country A (population 1 million) and 100,000 residents born in country B (population 10 million) has a stronger migratory pull to country A than B. National population totals of each country were obtained (United Nations Population Division 2008), and used to convert the migrant stock numbers into percentages of national origin country population.
RESULTS

Lack of recombination and strong phylogenetic signal. Likelihood mapping analyses were performed to quantify the nature of the phylogenetic information in the data. These indicated high phylogenetic signal (>85%) for all three datasets (Figure S1). In addition, the Thailand dataset, which includes 20 strains sampled during a 2006-2007 single hospital outbreak, was analyzed independently. Interestingly, the signal was the highest (98.8%) for the Thailand subset even though the sequences were only sampled over an interval of 154 days (Nickerson et al. 2009). The site frequency spectrum was calculated for the full (4310 SNP) dataset (Figure S2). While about 50% of the observed changes were singletons, about 15% of the changes were polymorphic at a frequency of >10%. This result is consistent with the strong phylogenetic signal revealed in the likelihood mapping analysis. Neighbor-nets were inferred for the full, synonymous and intergenic datasets (Figure S3). In all cases, no major splits were found, and the PHI test gave \( p \)-values greater than 0.001, suggesting low signal for recombination. This is consistent with a previous in vitro study that found limited uptake of foreign genetic material in \textit{S. aureus} (Waldron, Lindsay 2006).

HA-MRSA ST239 is measurably evolving population. Maximum likelihood trees were inferred for each of the three datasets using either a global clock model ("single rate"), or a clock model that incorporated times of sequence sampling ("single rate dated tips"). The two models were compared by taking twice the difference of the natural log of the likelihoods, which is assumed to follow a chi-squared distribution. In all three cases, the single rate dated tips model was a significantly better fit to the data than the single rates model (Table 1). This result indicates that ST239 HA-MRSA is a measurably evolving population.

Bayesian molecular clock analysis. The above results indicated that the HA-MRSA ST239 dataset is suitable for Bayesian phylogenetic analysis. Thus, Bayesian coalescent molecular clock analyses (Drummond, Rambaut 2007) were performed to reconstruct the temporal history of HA-MRSA ST239. Different molecular clock (strict vs. relaxed) (Drummond
et al. 2006a) and coalescent (constant vs. non-parametric growth; Heled, Drummond 2010) models were tested for the full, intergenic and synonymous datasets.

In any given alignment of closely related bacterial sequences, the vast majority of sites will be invariant, and thus phylogenetically uninformative. These sites can be excluded from the analysis as long as the site probabilities are renormalized using an ascertainment bias correction to account for the difference between unobserved and excluded site patterns. We report results with and without this correction (Table 2). Evidence against the null model of constant effective population size over time was only weak or moderate (Bayes factor BF < 10) (Suchard, Weiss, Sincheimer 2001b). Evidence against the simpler strict clock model was very strong (BF > 200) in all cases. Median estimates of the time to the most recent ancestor (TMRCA) of ST239 were similar across models, datasets, and with and without the bias correction, ranging from 1957 to 1972 with largely overlapping 95% highest posterior density (HPD) intervals (Figure 1), considerably earlier than the earliest identified ST239 strain from 1985. The consistency of timing among models reflected the strong temporal signal in the genome-wide SNPs data. The nucleotide divergence rates are given for each model in Table S2. In general, they are on the order of $10^{-3}$/site/year for the uncorrected estimates, and $10^{-4}$ for the ascertainment bias correction. For the full dataset, these rates can be transformed to reflect the divergence rate “per nucleotide site” (i.e. $[10^{-3} \times 4310]/[2.5 \times 10^6]$), resulting in rates of $10^{-6}$ to $10^{-7}$, consistent with recent estimates using related approaches (Harris et al. 2010; Nubel et al. 2010). However, the appropriate number of total sites in the denominator for the synonymous and intergenic sites is less obvious. It is clear that tight HPDs accompany these estimates, even under the relaxed molecular clock, suggesting that strong temporal information is present in the dataset. This result is in contrast to the predicted distribution of rates when little signal is present, i.e. wide confidence intervals that tend towards zero (Firth et al. 2010). Great caution should be used in interpreting these rates as long-term evolutionary or substitution rates, as it is likely they are closer to the mutation rate of the organism. Given that the estimates of the
TMRCAs were very similar among datasets with different rates, the information contained in the dataset is sufficient for inference of divergence times.

*Bayesian phylogeography of HA-MRSA ST239.* The spatiotemporal dynamics of HA-MRSA ST239 was reconstructed using a Bayesian phylogeographic model that assumed a relaxed molecular clock, constant population size, and included the ascertainment bias correction. Geographic coordinates were assigned to each sequence based on the cities from which the strains were sampled (see Methods). The full model, in which all 78 exchange rates are positive, was compared to a mixed model, in which a subset of the exchange rates is allowed to be zero through the use of a Bayesian stochastic search variable (BSSVS; Lemey et al. 2009). For all three priors (see Methods), under both the full and mixed models, the United States or Australia showed approximately equal probability for being the location for the origin of the sampled HA-MRSA ST239 strains, although their probabilities relative to the other locations were low (Figure 2). Given the small number of strains sampled in each location, and the absence of any strain phylogenetically close to the root, low confidence in assigning the location of the root is perhaps to be expected. Additional sequences would need to be included from each of the locations to ensure that sampled strains are representative. In the present study, the statistical power of each model to extract information from the data was assessed using the Kullback-Leibler (KL) divergence. In general, full models resulted in higher KL values than BSSVS models (Figure 2). This indicated that the posterior distributions of such models were in fact further away from the prior distributions (equal probability for all locations). The minimal rate configuration (see Methods) strongly favored by the mixed model is analogous to a parsimony model, which is often the default assumption used in many phylogeographic studies but rarely tested against other possibilities (Lemey et al. 2009). KL divergence was higher for the migration-informed prior (0.144) than the distance-informed (0.102) or flat (0.102) priors, under the full model, suggesting that human migration patterns are associated with the global migration of HA-MRSA ST239.
The maximum clade credibility (MCC) tree was obtained for the model with the highest KL divergence (full model with human migration prior, Figure 3). The geographic location of the internal nodes closest to the root had low posterior support (<0.5), suggesting that the early geographic spread of the epidemic, including the epidemic origin and initial routes of HA-MRSA ST239 spread, should be interpreted with caution. However, stronger support was found for the geographic reconstruction of the internal nodes in the more derived clades, particularly those from South America + Europe and Asia (Figure 3), which led to a robust reconstruction of bacterial gene-flow during the 1990s (Figure 4). Finally, the MCC tree was converted into a Google-Earth compatible movie (see Supplementary movie file) to visualize a possible scenario of ST239 global dissemination.

**DISCUSSION**

For the first time, we successfully employed Bayesian phylogeographic methods to a bacterial pathogen. *S. aureus* is a common cause of infections and therefore has not generally been thought of as a pandemic disease. However, with the emergence of specific antibiotic-resistant strains, it has become clear that *S. aureus* can spread in epidemic waves across the globe. In this study, we demonstrate that the TMRCA of the ST239 lineage occurred shortly after the introduction of Methicillin in 1959 (Batchelor et al. 1959). The type III SCCmec cassette also encodes resistance to penicillin, tetracycline, and erythromycin (Deurenberg, Stobberingh 2008), all of which were in use before 1960 (Khardori 2006). Multiple-drug resistance could have provided this strain with a strong fitness advantage in the hospital setting. We estimate that ST239 was spreading worldwide for almost twenty years before the initial identification of an isolate, in 1985, carrying the type III pathogenic cassette (Deurenberg, Stobberingh 2008). Such a discrepancy is alarming because it suggests that antibiotic-resistant bacteria can circulate globally undetected for a relatively long time. Except for a single introduction into Asia, ST239 lineages were not geographically restricted but rather regularly moved among
continents, indicating that HA-MRSA diffusion has the characteristics of a pandemic rather than regionally restricted outbreaks. Such results are in contrast with previous reports suggesting that MRSA is characterized by limited geographic dispersal in Europe (Grundmann et al. 2010) and worldwide (Nübel et al. 2008). However, these studies used limited genetic data, which may have contained too little resolution to infer detailed geographic patterns. Furthermore, because the genome-wide SNPs dataset used in the present analysis only included the core genome (excluding the SCC\textit{mec} cassette; Harris et al. 2010), the inferred evolutionary history is representative of the population history rather than the recombinant events within mobile elements.

The rate of evolution of bacteria species is uncertain (Achtman 2008). In this study, the use of the ascertainment correction bias provided estimates of evolutionary rates that were intermediate between those estimated using tip-dated sequences without the bias correction reported here and elsewhere (Harris et al. 2010; Nubel et al. 2010) and those inferred using species divergence dating techniques (Ochman, Elwyn, Moran 1999). It is likely that previously estimated rates without the bias correction were artificially high, as the lack of invariant site patterns are not taken into account, in which case the sum of all possible site pattern probabilities is <1. Correcting this bias allowed the invariant sites to be properly incorporated into the probability summation. Evolutionary rates estimates were still 100 times faster than rate estimates based on species divergence times (external calibrations). Reasons for this include the fact that transient polymorphisms exist within a population that will eventually be selected against over time, thereby driving up the estimated evolutionary rate when inferences are drawn from intra-population samples. Therefore, this rate reflects a different population process to the long-term fixation process between species. However, the important result from the present analysis is that estimates of bacterial evolutionary timescales are feasible using serially-sampled sequence data alone, and do not always require a calibration date or external rate.
 Recent theoretical advances in Bayesian phylogeography have made it possible to include the geographic location of the samples into the analysis, thus providing a formal statistical framework in which hypotheses concerning the spatial origin and dissemination of epidemics can be investigated. Bayesian methods account for phylogenetic uncertainty inherent in any reconstruction of the evolutionary history of a group of organisms, both in the tree topology and the assignment of geographic states to ancestral nodes, by estimating a probability distribution for parameters of interest (Lemey et al. 2009). This approach constitutes a significant improvement (Sanmartín, van der Mark, Ronquist 2008) upon traditional parsimony-based models (Slatkin, Maddison 1989) that only consider one reconstruction of migration on a fixed tree. Furthermore, while maximum parsimony can reliably be used in simple dissemination scenarios, minimizing the number of migrations is inappropriate in more complex situations, for example in cases of continuous multi-directional gene flow (Cunningham, Omland, Oakley 1998). The Bayesian framework, on the other hand, can account for more complex models by allowing the calculation of probabilities for the ancestral state (nucleotide and geographic location) reconstruction. This framework also allows testing of specific hypothesis of the driving factors of the migration (Lemey et al. 2009). Here, we show that human migration appears to be significantly associated with the current HA-MRSA pandemic spread. Travelling and migration have been linked anecdotally with community-acquired MRSA (CA-MRSA) (Ellington et al. 2009) and Methicillin-sensitive S. aureus (MSSA) (Schleucher, Gaessler, Knobloch 2008). Although CA-MRSA has traditionally been viewed as demographically and clinically distinct from HA-MRSA and more similar to MSSA (Groom et al. 2001; Naimi et al. 2003), recent evidence suggests that the epidemiological behavior of both CA- and HA-MRSA have begun to overlap (Seybold et al. 2006; Klevens et al. 2007). Furthermore, asymptomatic carriers of HA-MRSA could act as vectors in transmitting the pathogen (Zanger 2010); in fact, the majority of HA-MRSA cases in the U.S. from 2007 occurred outside of the hospital (Klevens et al. 2007), providing ample opportunity for transmission in the
community setting. Given the rapid increase of international air travel during the past several
decades, emergence of drug-resistant pathogens in a specific locale should not be considered
as an isolated event, but rather within a larger global context. Specific migration patterns (such
as those reported here) can then be incorporated into monitoring and intervention strategies.

To strengthen the phylogeographic analysis of viral gene flow, future studies should
include multiple strains from each sampled location to avoid uncertainty in the reconstruction of
ancestral locations. The cost of genome-wide typing is justified by the potential public health
utility of such data. Future applications of this approach may aid the control and prediction of
newly emergent drug-resistant pathogens.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>LnL$^c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>SR$^a$</td>
<td>-50848.15</td>
</tr>
<tr>
<td></td>
<td>SRDT$^b$</td>
<td>-50710.25*</td>
</tr>
<tr>
<td>Intergenic</td>
<td>SR</td>
<td>-7410.00</td>
</tr>
<tr>
<td></td>
<td>SRDT</td>
<td>-7365.51*</td>
</tr>
<tr>
<td>Synonymous</td>
<td>SR</td>
<td>-7718.34</td>
</tr>
<tr>
<td></td>
<td>SRDT</td>
<td>-7691.08*</td>
</tr>
</tbody>
</table>

1. **Table 1.** Phylogenetic test for measurably evolving populations.
2. $^a$ SR = single rate
3. $^b$ SRDT = single rate dated tips
4. $^c$ Ln likelihood for the specified model; values with an asterisk indicate significance for the SRDT over the SR, p<0.01.
<table>
<thead>
<tr>
<th>Dataseta</th>
<th>Clock Model</th>
<th>Coalescent Prior (CP)</th>
<th>Marginal Likelihood</th>
<th>Bayes Factor (CP)b</th>
<th>Bayes Factor (clock)c</th>
<th>Mean and 95% HPDs for the TMRCA of ST239</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>Strict</td>
<td>Constant</td>
<td>-32217.669</td>
<td></td>
<td></td>
<td>1958 1953 - 1962</td>
</tr>
<tr>
<td>F</td>
<td>Relaxed</td>
<td>Constant</td>
<td>-32094.852</td>
<td></td>
<td>245.634</td>
<td>1964 1949 - 1975</td>
</tr>
<tr>
<td>F</td>
<td>Relaxed</td>
<td>BSP</td>
<td>-32093.590</td>
<td>2.524</td>
<td>246.028</td>
<td>1968 1960 - 1975</td>
</tr>
<tr>
<td>S</td>
<td>Strict</td>
<td>Constant</td>
<td>-7530.590</td>
<td></td>
<td></td>
<td>1950 1940 - 1959</td>
</tr>
<tr>
<td>S</td>
<td>Relaxed</td>
<td>Constant</td>
<td>-7504.137</td>
<td></td>
<td>52.906</td>
<td>1954 1933 - 1974</td>
</tr>
<tr>
<td>F</td>
<td>Strict</td>
<td>Constant</td>
<td>-28238.886</td>
<td></td>
<td></td>
<td>1959 1955 - 1964</td>
</tr>
<tr>
<td>F</td>
<td>Strict</td>
<td>BSP</td>
<td>-28236.476</td>
<td>4.82</td>
<td></td>
<td>1959 1955 - 1964</td>
</tr>
<tr>
<td>F</td>
<td>Relaxed</td>
<td>Constant</td>
<td>-28119.184</td>
<td></td>
<td>239.404</td>
<td>1964 1950 - 1975</td>
</tr>
<tr>
<td>I</td>
<td>Relaxed</td>
<td>Constant</td>
<td>-6413.103</td>
<td></td>
<td>33.352</td>
<td>1972 1965 - 1978</td>
</tr>
<tr>
<td>S</td>
<td>Strict</td>
<td>Constant</td>
<td>-6538.188</td>
<td></td>
<td></td>
<td>1953 1944 - 1962</td>
</tr>
<tr>
<td>S</td>
<td>Relaxed</td>
<td>Constant</td>
<td>-6513.408</td>
<td></td>
<td>49.56</td>
<td>1957 1938 - 1975</td>
</tr>
<tr>
<td>S</td>
<td>Relaxed</td>
<td>BSP</td>
<td>-6509.456</td>
<td>7.904</td>
<td>46.994</td>
<td>1969 1951 - 1979</td>
</tr>
</tbody>
</table>
Table 2. Bayesian estimates of the evolutionary rate and median root for major clades.

The upper half of the table gives values without the ascertainment bias correction, and the lower half gives estimates incorporating the correction.

- Datasets are coded as F = full, I = intergenic, S = synonymous.
- Bayes Factors are given for the coalescent prior (CP): constant population size and the non-parametric Bayesian skyline plot (BSP) model (under the same clock model).
- Bayes Factors are given for the constant and relaxed molecular clock (under the same coalescent model)
Figure Legends.

Figure 1. Posterior estimates of the TMRCA for the ST239 phylogeny. Posterior estimates for 3 datasets (full, intergenic, and synonymous), two clock models (strict and relaxed), two coalescent priors (constant and Bayesian skyline plot) and with and without the ascertainment bias correction (ABC). The distributions are colored as follows: red = strict clock, constant; blue = strict clock, BSP, black = relaxed clock, constant, green = relaxed clock, BSP. Note that the earliest time is on the right of the x-axis, and the most recent times are on the left.

Figure 2. Bayesian posterior estimates of root location. Posterior probability estimates (y-axis) for the root location (x-axis) under the full and BSSVS models with three different priors (flat, distance-informed and migration-informed). The Kullback-Leibler (KL) divergence measure is reported for each model/prior. The model with the largest KL is highlighted in bold. All analyses assumed a relaxed clock, GTR model of nucleotide substitution, and constant coalescent prior.

Figure 3. Bayesian inference of ST239 evolutionary history. Maximum clade credibility (MCC) tree selected from the posterior distribution for the model/prior with the highest KL divergence. Branches are scaled by time according to the scale at the bottom. Terminal branches are colored by location of sampling according to the legend on the left. Internal branches are colored according to the most probable location. Posterior support for the geographic location for each internal node is given along the subtending branch. All nodes that are unlabeled have a posterior >0.90. Nodes with low posterior probabilities are labeled A-F. Piecharts are shown for these nodes at the bottom of the figure, with the support for each geographic location shown as a proportion of the circle, colored according to the same color scheme as the tree. All analyses assumed a relaxed clock, GTR model of nucleotide substitution, and constant coalescent prior.

Figure 4. Spatial spread of HA-MRSA ST239 during the 1990s. Major events in the global dissemination of HA-MRSA are shown for (A) Europe and South America and (B) Asia.
corresponding to the South America + Europe and Asia clades in Figure 3. Countries are colored according the legend. Lines denote migration routes between countries with the arrow indicating the direction of the migration.


