MICROBIAL GENOMICS
Bases to Biology

Research Paper

Genome-scale rates of evolutionary change in bacteria

Sebastian Duchêne,[1,2,3] Kathryn E. Holt,[2,3] François-Xavier Weill,[4] Simon Le Hello,[4] Jane Hawkey,[2,3] David J. Edwards,[2,3] Mathieu Fourment[1] and Edward C. Holmes[1]

[1]Marie Bashir Institute of Infectious Diseases and Biosecurity, Charles Perkins Centre, School of Life and Environmental Sciences and Sydney Medical School, The University of Sydney, Sydney, NSW 2006, Australia

[2]Centre for Systems Genomics, The University of Melbourne, Melbourne, VIC 3010, Australia

[3]Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, The University of Melbourne, Melbourne, VIC 3010, Australia

[4]Institut Pasteur, Unité des Bactéries Pathogènes Entériques, Paris 75015, France

Correspondence: Edward C. Holmes (edward.holmes@sydney.edu.au)

DOI: 10.1099/mgen.0.000094

Estimating the rates at which bacterial genomes evolve is critical to understanding major evolutionary and ecological processes such as disease emergence, long-term host–pathogen associations and short-term transmission patterns. The surge in bacterial genomic data sets provides a new opportunity to estimate these rates and reveal the factors that shape bacterial evolutionary dynamics. For many organisms estimates of evolutionary rate display an inverse association with the time-scale over which the data are sampled. However, this relationship remains unexplored in bacteria due to the difficulty in estimating genome-wide evolutionary rates, which are impacted by the extent of temporal structure in the data and the prevalence of recombination. We collected 36 whole genome sequence data sets from 16 species of bacterial pathogens to systematically estimate and compare their evolutionary rates and assess the extent of temporal structure in the absence of recombination. The majority (28/36) of data sets possessed sufficient clock-like structure to robustly estimate evolutionary rates. However, in some species reliable estimates were not possible even with 'ancient DNA' data sampled over many centuries, suggesting that they evolve very slowly or that they display extensive rate variation among lineages. The robustly estimated evolutionary rates spanned several orders of magnitude, from approximately 10^{-5} to 10^{-8} nucleotide substitutions per site year^{-1}. This variation was negatively associated with sampling time, with this relationship best described by an exponential decay curve. To avoid potential estimation biases, such time-dependency should be considered when inferring evolutionary time-scales in bacteria.

Keywords: evolution; bacteria; phylogeny; substitution rates; time-dependency; molecular clock.

Abbreviations: HPD, 95% highest posterior density; ML, maximum-likelihood.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files.

Data Summary

This study consists of nucleotide sequence alignments, all of which are available online (http://zenodo.org/record/45951#.Vr1M-JN95E4).

Received 16 September 2016; Accepted 24 October 2016

Introduction

Estimating the rate of molecular evolution is critical for understanding a variety of evolutionary and epidemiological processes. Rates of molecular evolution are the product of the number of mutations that arise per replication event, the frequency of replication events per unit time and the probability of mutational fixation. These rates are determined by a variety of factors including background mutation rate, the direction and strength of natural selection, generation time and

population size. Both generation time and population size have been shown to scale negatively with the substitution rate in a range of organisms (Bromham, 2009). For example, because of differences in generation time, spore-forming bacteria evolve more slowly over time than those that do not form spores (Weller & Wu, 2015). Similarly, lineages undergoing adaptive evolution are expected to accumulate substitutions more rapidly than those subject to purifying selection (Eyre-Walker & Keightley, 2007).

Despite the wealth of sequence data, genome-wide rates of evolutionary change in bacteria are often uncertain. At one end of the spectrum, rates as high as $\sim 10^{-5}$ nucleotide substitutions per site year^{-1} have been reported for *Neisseria gonorrhoeae* (Pérez-Losada *et al.*, 2007). In contrast, genome-wide rates of only $\sim 10^{-9}$ substitutions per site year^{-1} have been observed in *Mycobacterium tuberculosis* (Comas *et al.*, 2013). Importantly, however, these estimates are not always readily comparable because they use different methods and sources of data. Furthermore, most previous studies have not investigated the degree of temporal structure (i.e. clock-like behaviour) in the data, such that their reliability is uncertain. The increasing availability of genomic data means that it is now possible to estimate substitution rates with sequences collected over a number of years (i.e. tip-date calibration), not only for rapidly evolving pathogens, such as RNA viruses, but also in some DNA viruses and bacteria (Biek *et al.*, 2015). The rates at which bacteria evolve, the strength of clock-like signal in the data and the determinants of any rate differences observed have not been systematically investigated. However, these parameters are of importance both for the accurate interpretation of outbreak investigations that may depend on the reliable estimation of the time-scale of putatively linked transmission cases, and for revealing the long-term time-scales over which bacteria have been associated with specific hosts.

Importantly, estimates of evolutionary rate have been shown to scale negatively with their time-scale of measurement in several organisms (Ho *et al.*, 2011), a pattern that has been attributed to the gradual purging of deleterious mutations over time (Ho & Larson, 2006; Penny, 2005). In the context of phylogenetic analyses, the time-scale of measurement corresponds to the age of the calibration or the sampling time-frame (Duchêne *et al.*, 2014; Molak & Ho, 2015). In viruses, natural selection, mutational saturation and substitution model inadequacy have also been shown to contribute to this pattern (Duchêne *et al.*, 2014, 2015a; Ho *et al.*, 2015a). However, the phenomenon of time-dependency of rate estimates remains largely unexplored in bacterial genomes, although there is some empirical evidence that it may also hold in these organisms (Comas *et al.*, 2013).

To provide a comprehensive picture of genomic-scale evolutionary rates in bacteria and their temporal dynamics, particularly the extent of time-dependency in the data, we analysed, using a variety of phylogenetic methods, 36 publically available whole genome SNP data sets from bacterial pathogens associated with human disease sampled over periods extending over 2000 years.

Impact Statement

Phylogenetic methods can be used to infer the evolutionary time-scale of groups of organisms. In some pathogens it is possible to estimate the time of origin of disease outbreaks or of cross-species transmission events. However, the accuracy of these estimates relies on our understanding of the rate at which genetic change accumulates over time. The recent surge of genomic data presents an unprecedented opportunity to enhance our understanding of microbial evolution, with the potential of improving future inferences of their evolutionary time-scale. We estimated the rate of nucleotide substitution in 36 complete genomes of different bacterial pathogens using a range of computational methods. We find large differences in the rates produced. For example, some bacteria, such as those that cause hospital-derived infections, evolve orders of magnitude more rapidly than those that cause tuberculosis, which undergo extended periods of latency. We also find that the sampling times appear to play an important role in determining their rate. Our results provide the first genomic perspective of bacterial rates of evolution, thereby improving our understanding of the time-scale over which they diversify.

Methods

Data collection. We collected 35 whole genome nucleotide SNP alignments from previously published studies (Baines *et al.*, 2015; Bart *et al.*, 2014; Bos *et al.*, 2014; Bratcher *et al.*, 2014; Croucher *et al.*, 2011; Davies *et al.*, 2015; Devault *et al.*, 2014; Eldholm *et al.*, 2015; Gaiarsa *et al.*, 2015; Holden *et al.*, 2013; Holt *et al.*, 2008, 2013, 2016; Howden *et al.*, 2013; Marvig *et al.*, 2013; Merker *et al.*, 2015; Njamkepo *et al.*, 2016; Schuenemann *et al.*, 2013; Schultz *et al.*, 2016; Stinear *et al.*, 2014; Uhlemann *et al.*, 2014; Wagner *et al.*, 2014; Ward *et al.*, 2014; Zhou *et al.*, 2013, 2014), and one unpublished data set of *Salmonella enterica* serovar Kentucky (Table S1). The *Salmonella enterica* serovar Kentucky data set comprised 88 strains isolated between 1937 and 2012 (Le Hello *et al.*, 2013). Whole genome sequencing was performed at GATC Biotech (Germany) using an Illumina HiSeq system and analysed as described previously (Holt *et al.*, 2013). Although sequence reads for most other data sets are publically available in the NCBI Short Read Archive (SRA), we obtained the original alignments from the authors wherever possible. With this approach we take advantage of domain knowledge of the original studies for choice of reference sequence and identification of repetitive or horizontally transferred sequences, which are important for the accurate generation of SNP

alignments. We removed outgroup taxa and samples that were distantly related to the majority of sequences to limit our analyses to the taxonomic group of interest and to minimize the artificial inflation of among-lineage rate variation due to the presence of very long branches in the phylogenetic trees. For the *Neisseria meningitidis* and *M. tuberculosis* Lineages 2 and 4 data sets we obtained sequence reads from the Short Read Archive and called SNPs using the RedDog pipeline [as described previously (Schultz *et al.*, 2016); available at https://github.com/katholt/RedDog].

SNPs can be introduced into bacterial genomes individually via mutations or in clusters by recombination, a process that may bias demographic inferences and rate estimates (Hedge & Wilson, 2014; Lapierre *et al.*, 2016). In particular, ignoring recombination can lead to incorrect estimates of branch lengths, which in turn results in an apparently overdispersed molecular clock, precluding accurate estimates of evolutionary rates and time-scales. In some cases, removing sites with evidence of recombination can alleviate this problem (see Fig. S1). Thus, we removed all genomic regions with evidence of recombination using Gubbins v1.4.2 (Croucher *et al.*, 2014), and verified these results using RDP4 (Martin *et al.*, 2015). For RDP4 we removed sequences with significant evidence of recombination according to six of the methods implemented in the program: Bootscan (Salminen *et al.*, 1995), Geneconv (Padidam *et al.*, 1999), Maxchi (Smith, 1992), Siscan (Gibbs *et al.*, 2000), RDP (Martin & Rybicki, 2000) and 3seq (Boni *et al.*, 2007). Accordingly, we discarded an entire data set of *N. meningitidis* in which nearly half of the sequences displayed recombination (Table S1). We also visually checked that the alignments were free of additional obviously recombining regions. Our final data set comprised 16 bacterial species from 13 genera, with genome sizes ranging from 1.4 Mbp (*Streptococcus pyogenes*) to 6.2 Mbp (*Pseudomonas aeruginosa*). The data set sizes ranged from 189 (*Streptococcus pneumonia* and *M. tuberculosis* Lineage 4) to 15 (*Mycobacterium leprae* ancient DNA) sequences, and alignment lengths from 15 394 SNPs (*Acinetobacter baumannii*) to 402 (*M. tuberculosis* Lineage 4) SNPs.

All the data sets analysed here included sampling times associated with each sequence in the form of year of isolation. Individual data sets spanned sampling ranges of approximately 2000 years in the case of *M. leprae* to 2.5 years in *Staphylococcus aureus* ST8:USA300. Four data sets included ancient DNA samples: *M. leprae* (Schuenemann *et al.*, 2013), two *Yersinia pestis* data sets (Wagner *et al.*, 2014), and one *M. tuberculosis* data set comprising strains sampled from New World mummies, animal strains and human Lineage 6 (Bos *et al.*, 2014). In *M. leprae*, the oldest sample had an estimated age of approximately 2000 years (Schuenemann *et al.*, 2013), while in the *Y. pestis* data sets the oldest samples had ages of approximately 600 and 1500 years (Wagner *et al.*, 2014). The New World mummy *M. tuberculosis* samples had ages of approximately 900 years (Bos *et al.*, 2014). Notably, one of the *Y. pestis* data sets included samples from both contemporary strains and those from the first, second and third plague pandemics, while the other is a subset that only included sequences from the second pandemic that began with the Black Death and contemporary strains. The remaining data sets had sampling times from a few years to several decades (Table S1). All the data sets used here are available online (http://zenodo.org/record/45951#.Vr1M-JN95E4).

Maximum-likelihood (ML) analysis and root-to-tip regression. We estimated ML trees using PhyML v3.1 (Guindon *et al.*, 2010), employing the GTR+Γ substitution model with four categories for the Γ distribution of among-site rate heterogeneity and sub-tree pruning regrafting branch-swapping. We did not consider the proportion of invariable sites in the substitution model because the data sets comprised SNPs only, such that all sites are variable. The branch lengths in the ML trees are the expected number of nucleotide substitutions per site; as we are working with SNP alignments, this corresponds to the expected number of substitutions per variable site. To convert the branch lengths into genome-wide distances, we multiplied them by the length of the SNP alignment, and divided them by the core genome length that reflects the size of the genome in which SNPs were called (this information was extracted from the original publications). We fitted regressions for the root-to-tip distance as a function of sampling time using NELSI v0.21 (Ho *et al.*, 2015b), where the position of the root is selected to maximize the determination coefficient, R^2. Under clock-like evolution there should be a linear relationship between sampling time (year) and the expected number of nucleotide substitutions along the tree. The slope of the line corresponds to the substitution rate, although this estimate is statistically invalid because data points may not be phylogenetically independent. The extent to which the points deviate from the regression line reflects the amount of among-lineage rate variation, measured using R^2 (Rambaut *et al.*, 2016).

Bayesian analysis and date randomizations. We estimated rates of evolutionary change using a Bayesian Markov chain Monte Carlo method implemented in BEAST v1.8 (Drummond *et al.*, 2012), with a chain length of 10^8 steps, sampling every 5000 steps. If the effective sample size of any of the parameters was less than 200, we increased the chain length by 50 % and reduced the sampling frequency accordingly. For all data sets we used the GTR+Γ substitution model and the sampling times (tip dates) for calibration. For this analysis we utilized both strict and uncorrelated lognormal molecular clock models and constant-size coalescent and Bayesian Skyline demographic models, resulting in four possible model combinations. We compared the statistical fit of these models by estimating marginal likelihoods via path-sampling (Xie *et al.*, 2011). We report the rate estimates from the model with the highest marginal likelihood.

To validate our Bayesian rate estimates we conducted a date-randomization test, which consists of repeating the analysis while assigning the sampling times randomly to the sequences

(Firth *et al.*, 2010). In all cases we used the best-fit combination of demographic and clock model, described above. We conducted ten randomization replicates per data set to generate an expectation of the rate estimates in the absence of temporal structure, which appears to be sufficient to assess temporal structure in bacterial data (Murray *et al.*, 2015). Two criteria have been proposed to assess temporal structure. One method, known as CR1 (Duchêne *et al.*, 2015b), considers that data do not have temporal structure if the mean rate with the correct sampling times is contained within the 95 % highest posterior density (HPD) of that from any of the randomizations. CR2 is more conservative; the data do not have temporal structure if the HPD of the estimate with the correct sampling times overlaps with that from any of the randomizations. These criteria have different levels of type I and type II errors (Duchêne *et al.*, 2015b). Here, we considered the proportion of randomized replicates with HPDs that overlapped with those obtained using the correct sampling times, following CR2. We arbitrarily determined that data had 'strong' temporal structure if the proportion was 0, 'moderate' if it was between 0 and 0.5, and 'low' if it was less than 0.5. We verified that samples with the same sampling time did not form monophyletic groups in our ML trees. Such pattern can produce false positives in the date-randomization test (i.e. incorrectly suggesting that a data set has temporal structure) (Duchêne *et al.*, 2015b; Murray *et al.*, 2015).

Regression analyses of the time-dependency of substitution rates. A key element of our study was to determine whether the time-span of sampling was associated with the evolutionary rate estimate. To test for this pattern we conducted a least squares linear regression for the genome-wide rate as a function of sampling time. Importantly, only rate estimates with strong and moderate temporal structure were included in this analysis. We used a $\log_{10}$ transformation because our rate estimates span several orders of magnitude. A potential shortcoming of this analysis is that the data do not necessarily represent independent samples because some correspond to closely related lineages. This can be addressed by using phylogenetic independent contrasts, or phylogenetic generalized least squares. However, these methods require a phylogenetic tree of the evolutionary relationships of all data points that cannot be estimated for our data because the sites in different SNP alignments are not necessarily homologous. Instead, we used an approach in which the regression is conducted using a single randomly chosen data point from each species. We repeated this procedure 1000 times and verified that the range of slope estimates with random subsamples did not include zero, and we report the mean value and the corresponding confidence interval. For our regression of the rate as a function of sampling time we used a test described previously (Duchêne *et al.*, 2014) to verify that the slope estimate was not a statistical artefact that sometimes occurs when fitting a regression for a ratio as a function of its denominator.

The linear regression method, however, does not provide a realistic description of time-dependency, which has been suggested to follow a decay curve (Ho *et al.*, 2011; Penny, 2005). We therefore modelled the relationship between rate and time using a double exponential curve of the form $r=a/T^b$, where r and T correspond to the rate estimate and sampling time on $\log_{10}$ scale, respectively, and parameters a and b control the asymptote and rate of decay in the function. This parameterization is similar to that proposed by O'Fallon (2010). Parameters a and b were optimized using the algorithm of Nelder & Mead (1965). To obtain a confidence interval around the parameter estimates, we conducted 100 bootstrap replicates of the rate estimates and sampling times.

Results

Measuring the extent of clock-like structure in bacterial evolution

Our root-to-tip regressions revealed large differences in the degree of clock-like behaviour, with 22 of the 35 data sets having R^2 values of less than 0.5, suggesting weak clock-like behaviour, and eight with R^2 values of 0.7 or higher, suggesting stronger clock-like behaviour (Fig. 1). A data set of *Staphylococcus aureus* multilocus sequence type ST239 had the highest R^2, at 0.96, with similar values observed in *Vibrio cholerae* (0.92), *Shigella sonnei* (0.89) and *Shigella dysenteriae* type 1 (0.88). The lowest R^2 was 1.52×10^{-3} for *M. tuberculosis* Lineage 2. The regression slope (rate) for *M. tuberculosis* Lineage 2 and *Y. pestis* from the second global pandemic included negative values, suggesting that these rates are either too low, or that there is extensive among-lineage rate variation, to allow reliable rate estimation.

Our comparison of molecular clock models included the strict clock, which assumes that the rate of evolution is constant across lineages, and the uncorrelated lognormal relaxed clock that treats the rate across lineages as a random variable. Most bacterial data sets favoured a relaxed molecular clock, such that there is marked rate variation among lineages (Figs S2 and S3). The date-randomization test suggested that 28 data sets had strong to moderate temporal signal (Figs 2 and S2) (defined as $\leq$5 randomizations with an HPD overlapping with the HPDs estimated with the correct tip-dates). Only one (*M. leprae)* of the 15 bacterial species (excluding *N. meningitidis*, which was discarded because of the large number of recombinant sequences) analysed had no data sets displaying moderate to strong temporal signal. However, five of eight species represented by two or more data sets displayed a mix of both weak signal and strong to moderate temporal signal (Fig. 2), suggesting that a lack of temporal signal may be a property of the individual data sets rather than a true species effect.

That even data sets from slowly evolving bacteria such as *M. tuberculosis*, which exhibit substitution rates in the order of 10^{-8} substitutions per site year^{-1}, possessed moderate temporal signal underlines the power of genome-scale data

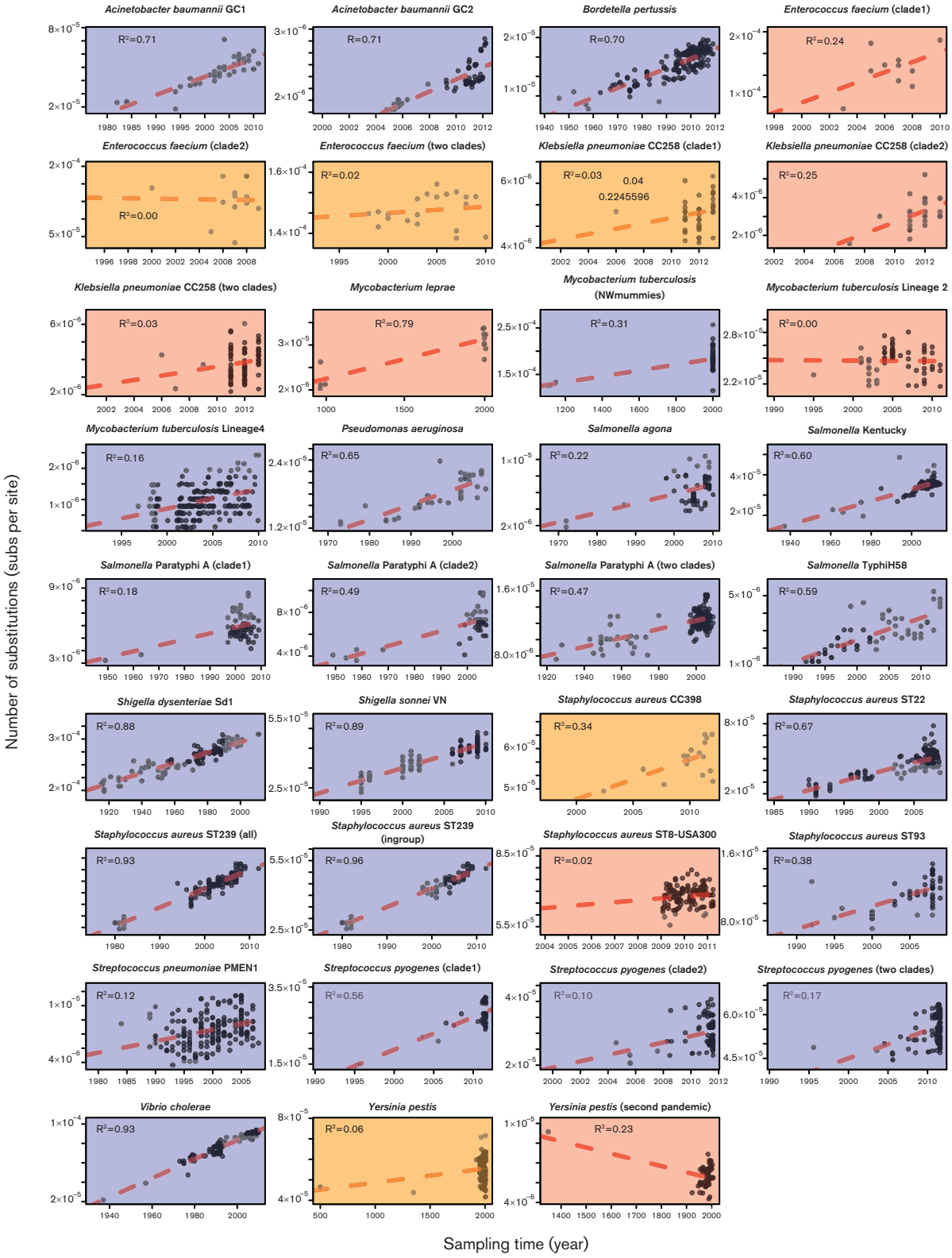

Fig. 1. Regressions of the root-to-tip genetic distance (expected nucleotide substitutions per site) as a function of sampling time (year) for 35 bacterial data sets. Each point corresponds to an individual sampled genome (SNP sequence in the alignment), and the red dashed line is the linear regression using least squares; the R^2 coefficients are also shown. Shading corresponds to the degree of temporal structure according to the date-randomization test in BEAST; blue indicates strong temporal structure, while orange and red indicate moderate and low temporal structure, respectively.

to estimate evolutionary dynamics. Despite this, it is striking that some data sets, including ancient DNA samples, had very little temporal signal, such that incorporating historical DNA data is not always sufficient for reliable rate estimation. For example, previous studies have suggested no temporal structure in data sets of *Y. pestis* with sampling time ranges of ~290 years (Bos *et al.*, 2016) and ~650 years (Wagner *et al.*, 2014), and of *M. leprae* with a sampling range of almost 2000 years (Schuenemann *et al.*, 2013).

The range of genome-wide rates of evolutionary change in bacteria

The highest mean genome-wide rate estimate for the data sets with temporal structure was 9.35×10^{-6} substitutions per site year^{-1} (HPD: 2.50×10^{-6}–1.74×10^{-5}) for a vancomycin-resistant *Enterococcus faecium* (VRE) data set sampled over 10 years. High rates were also observed in *A. baumannii* Global Clone 2 (GC2) sampled over 7 years and with a mean rate of 3.15×10^{-6} substitutions per site year^{-1} (HPD: 2.34×10^{-6}–4.44×10^{-6}) and *Staphylococcus aureus* clonal complex 398 (CC398) sampled over 9.5 years and with a mean rate of 2.43×10^{-6} substitutions per site year^{-1} (HPD: 1.14×10^{-6}–3.98×10^{-6}). The lowest estimate was for *Y. pestis* with samples collected over ~1500 years which resulted in a mean rate of 1.57×10^{-8} substitutions per site year^{-1} (HPD: 1.03×10^{-8}–2.27×10^{-8}), although this data set had only moderate temporal structure. Other data sets with low rates and strong temporal structure were: *M. tuberculosis*, with samples collected over ~900 years and a mean rate of 5.39×10^{-8} substitutions per site year^{-1} (HPD: 2.49×10^{-8}–1.02×10^{-7}); *M. tuberculosis* Lineage 4 with samples collected over 13 years and a mean rate of 5.67×10^{-8} substitutions per site year^{-1} (HPD: 3.80×10^{-8}–8.02×10^{-8}); and three *Salmonella enterica* serovar Paratyphi A data sets, with sampling times of 58–84 years and mean rates ranging from 7.60×10^{-8} to 9.47×10^{-8} substitutions per site year^{-1}(Figs 2 and S2).

There was large variation in rate estimates for some closely related bacteria. Our analyses included several data sets of different serovars of *Salmonella enterica*: Kentucky, Agona, Typhi and Paratyphi A. Interestingly, the rate estimates for the human-restricted serovars Typhi and Paratyphi A (the agents of typhoid fever) (mean rates ranging from 1.78×10^{-7} to 8.02×10^{-8} substitutions per site year^{-1}) were consistently lower than those estimated for the host-generalist serovars Kentucky and Agona (5.34–3.95×10^{-7} substitutions per site year^{-1}). The rate estimates for *Staphylococcus aureus* were also highly variable between lineages, which included CC398, type USA300 and multilocus sequence types ST22, ST93 and ST239. The estimated mean rate for the livestock-associated CC398 was 2.43×10^{-6} substitutions per site year^{-1} (HPD: 1.14×10^{-6}–3.86×10^{-6}), the highest for this species. In contrast, the estimated mean rate for ST93 was 5.55×10^{-7} substitutions per site year^{-1} (HPD: 2.77×10^{-7}–9.60×10^{-7}), nearly five times lower (Fig. 2).

We also investigated whether rate estimates based on root-to-tip regression were consistently biased compared to those obtained using the Bayesian approach in BEAST (Fig. 3). If the estimates from the two methods are the same, then they should fall along the line $y=x$ when plotted against each other. Notably, most points fell above the regression line, implying that the mean Bayesian estimates (y-axis) were higher than those from the regression (x-axis). A probable cause of this pattern is that deep branches in the phylogenetic trees are over-represented in the regression method. If substitution rates do indeed vary in a time-dependent manner (Duchêne *et al.*, 2015a; see below), such that higher rates are observed toward the present, then rate estimates obtained using regression may exhibit a downward bias (Fig. 3). The exceptionswere three *Klebsiella pneumoniae* data sets (CC258), only one of which had sufficient temporal structure for reliable estimation, with a mean rate estimate using BEAST of 2.99×10^{-7} substitutions per site year^{-1} (HPD: 9.51×10^{-8}–5.65×10^{-7}), while that using regression was 3.05×10^{-7} substitutions per site year^{-1} [95 % confidence interval (CI): 1.03×10^{-7}–5.07×10^{-7}); and *Bordetella pertussis*, with a mean rate of 1.74×10^{-7} substitutions per site year^{-1} using BEAST (HPD: 1.53×10^{-7}–1.94×10^{-7}), and of 2.15×10^{-7} substitutions per site year^{-1} (CI: 2.24×10^{-7}–2.78×10^{-7}) using regression.

Modelling time-dependent rates of evolution

For those data sets with strong to moderate temporal structure, we investigated the relationship between evolutionary rate and sampling time, considered as the time-span between the youngest and oldest samples in each data set. Our linear regression for the rate estimates as a function of sampling time on a $\log_{10}$ scale revealed a significant negative association between rate and time with slope $=-0.701$ (CI: -4.74 to -5.81 and $P=0.0003$; Fig. 4). The optimal parameterization for the decay function $r=a/T^b$ was $a=-5.90$ (CI: -6.39 to -5.77) and $b=-0.17$ (CI: -0.26 to -0.13) (Fig. 4). Importantly, the residual errors were normally distributed around the fitted curve (Fig S4).

Discussion

We have performed the largest comparative and systematic study of evolutionary rates in bacteria to date, providing an important resource for future studies of bacterial evolution. This analysis revealed an approximately two order of magnitude range of evolutionary rates for those bacterial data sets in which there was sufficient temporal structure for rate estimation. For the most rapidly evolving bacteria analysed here, such as *E. faecium*, *Staphylococcus aureus* and *A. baumannii*, we estimated genome-wide rates between 10^{-5} and 10^{-6} substitutions per site year^{-1}. Not only are these rates within the range of those previously estimated for rapidly evolving bacteria (Holt *et al.*, 2012; Ward *et al.*, 2014) but, more strikingly, they are also similar to those of slowly evolving DNA viruses (Duchêne *et al.*, 2014; Firth *et al.*, 2010). However, even our highest rate estimates are lower than some reported previously, such as those for

Fig. 2. Bayesian estimates of genome-scale nucleotide substitution rates for all bacterial data sets. The axis for the nucleotide substitution rate is shown on $\log_{10}$ scale. Circular points represent the mean rate estimate, and error bars correspond to the 95 % HPD values. Colours indicate the degree of temporal structure according to the date-randomization test as indicated in Fig. 1. For comparison, the x symbol represents the point estimates using regression, while the asterisk (*) corresponds to estimates that were negative, and are thus not shown.

Helicobacter pylori (Kennemann *et al.*, 2011) and *N. gonorrhoeae* (Pérez-Losada *et al.*, 2007), at ~10^{-5} and 10^{-4} substitutions per site year^{-1}, respectively. As these species also experience high rates of recombination (Maixner *et al.*, 2016; Yahara *et al.*, 2016), which can disrupt temporal signal, we suggest that these unusually high estimates be treated with caution until they are verified. For example, our *N. meningitidis* data set exhibited extensive recombination, precluding standard phylogenetic analyses.

The lowest rate estimates obtained here, at ~10^{-8} substitutions per site year^{-1}, were also comparable to previous studies of slowly evolving bacteria, notably *M. tuberculosis* (Kay *et al.*, 2015). Importantly, however, even though they were relatively low, these rates were sufficiently rapid to be accurately estimated using genomic-scale data. In contrast, rate estimates lower than about 1.5×10^{-8} substitutions per site year^{-1} could not be validated using our methods and hence may require data sampled over far longer time periods. However, it was also noteworthy that the *M. leprae* and one of the *Y. pestis* data sets did not have sufficient temporal structure for rate estimation despite the availability of ancient DNA. Interestingly, a recent analysis of Bronze Age strains of *Y. pestis* is consistent with a much higher substitution rate in this bacterium, at ~10^{-7} substitutions per site year^{-1} (Rasmussen *et al.*, 2015), in contrast to the data presented here and previously (Cui *et al.*, 2013; Morelli *et al.*, 2010). Why these data sets differ so fundamentally in estimated substitution rate is clearly an important area for future study.

Nearly all bacterial species investigated here displayed genome evolution that was measurable over a period of 10–100 years. Data sets with sampling times spanning less than

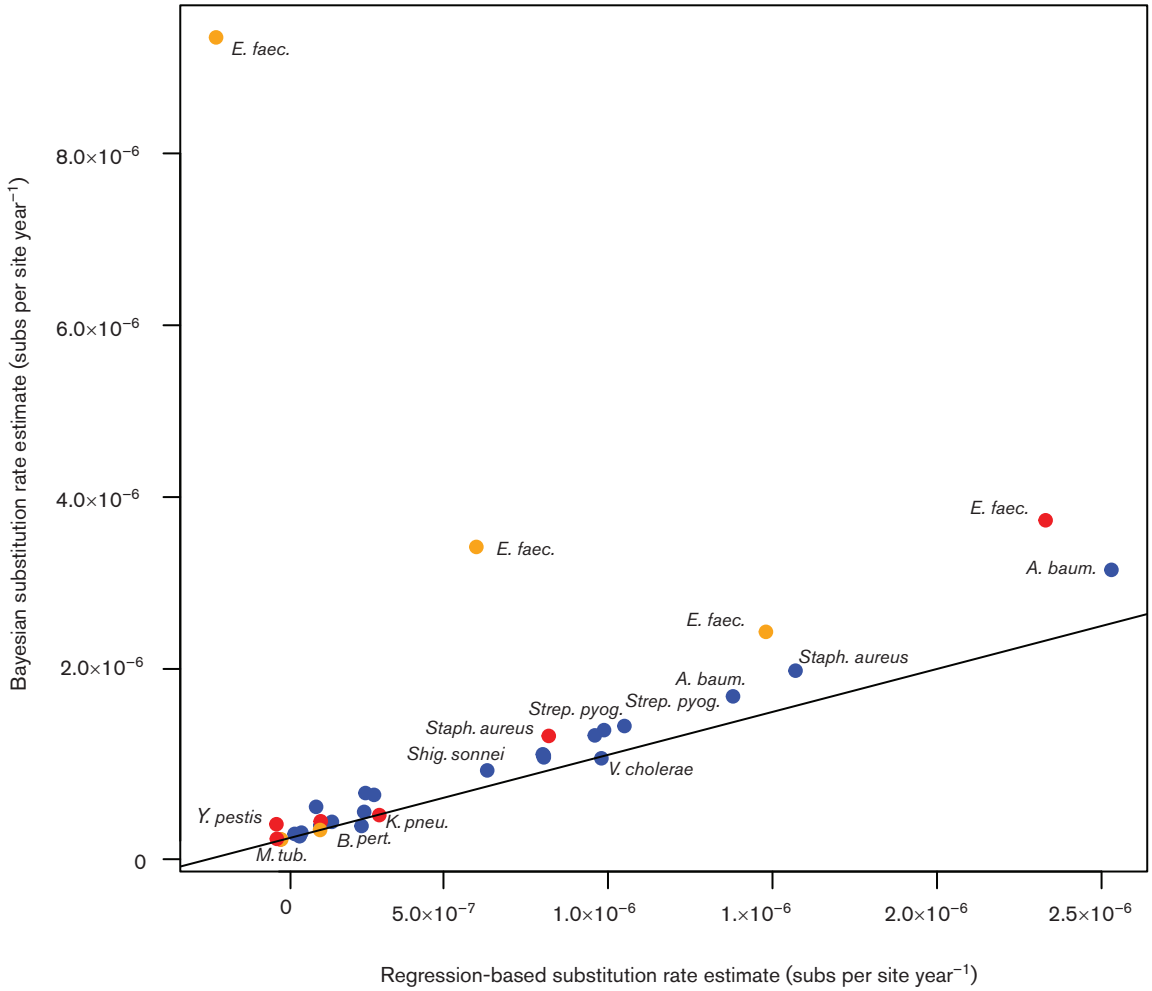

Fig. 3. Estimates of genome-scale nucleotide substitution rates using a Bayesian method (BEAST) compared to those estimated via root-to-tip regression. The line represents $y=x$. Points that fall on the line correspond to data sets for which the mean Bayesian estimates closely match those from the regression. Points above and below the line are data sets for which the Bayesian estimate is higher or lower than that from the regression, respectively. The colour corresponds to the degree of temporal structure according to the date-randomization test as indicated in Fig. 1.

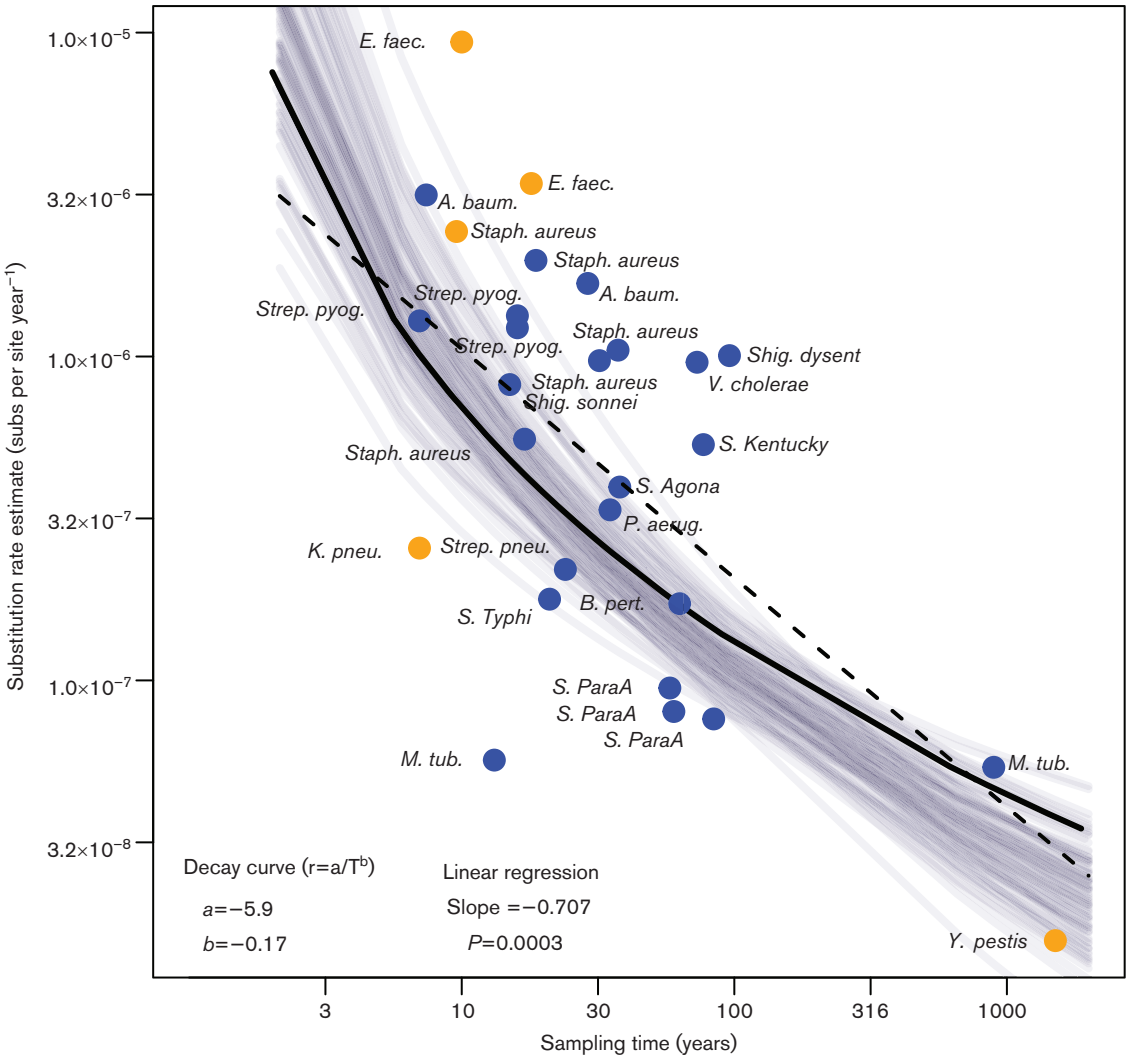

Fig. 4. Estimates of genome-wide nucleotide substitution rates in human-associated bacterial pathogens as a function of sampling time in years. The axes are shown on a $\log_{10}$ scale. The colour corresponds to the degree of temporal structure according to the date-randomization test; blue indicates strong temporal structure and orange indicates moderate temporal structure. The dashed line corresponds to the linear regression, while the solid line corresponds to the decay curve (both fitted using only the points with strong and moderate temporal structure). The grey lines represent 100 bootstrap replicates of the decay curve, and thus represent the uncertainty in the decay function.

10 years were largely unreliable. Importantly, for several species the strength of temporal signal varied substantially between data sets, and a lack of temporal signal in one data set could not be taken as a general indication of overall lack of signal for the species.

Notably, our analysis reveals that evolutionary rates in bacteria display a negative relationship with sampling time, such that their rates can be considered time-dependent (Ho *et al.*, 2011). This observation is of particular importance in the context of the analysis of genome sequences taken from individual disease outbreaks or transmission chains (Didelot *et al.*, 2012), as the evolutionary rates measured over these very short time-scales will probably contain deleterious mutations yet to be purged by purifying selection and substitutional saturation is increasingly apparent over time. As such, these estimates are expected to be much higher than those for samples obtained over longer time-scales. For example, based on our data (Fig. 4), we would predict the evolutionary rate estimated for a given bacterial species or clade over a sampling frame of 10 years to be more than an order of magnitude higher than that estimated for the same bacteria sampled over a period of 100 years. Conversely, the longer-term (and lower) evolutionary rates of the type inferred here would tend to over-estimate the time-scale of bacterial transmission when applied to outbreak data. The relationship between evolutionary rate and sampling time can be used to assess the extent to which extrapolating rates of evolution over different time-scales will lead to a bias in estimates of divergence times. To this end, future studies should compare bacterial data sets of the same species

across different sampling time-frames to more precisely determine the nature of the time-dependent curve.

Strikingly, in some cases the time-dependent pattern appears to hold within closely related bacterial lineages, such as our *A. baumanii* and *Salmonella enterica* Paratyphi A data sets. However, it is notable that both reliable rate estimates for *M. tuberculosis*, estimated over sampling frames of 15 and 895 years, were nearly identical. *M. tuberculosis* are notoriously slow growing bacteria whose generation time is orders of magnitude slower than the other bacteria analysed here. It is likely that other factors such as genome size and DNA G+C (guanine and cytosine) content also contribute to rate variation between bacteria (Rocha *et al.*, 2006). Given the time dependency of rate estimates we observed, we suggest future studies of bacterial evolutionary dynamics would be best addressed by comparing multiple independent replicate sample sets for each species, collected over matched time-spans. Overall, our analyses show that genome evolution can now be reliably measured in bacteria and establish a genomic framework for understanding long-term evolutionary dynamics in bacteria.

Acknowledgements

This research was funded by the NHMRC (Australia Fellowship #AF30 to E. C. H., Career Development Fellowship #1061409 to K. E. H.). S. D. was supported by a McKenzie fellowship from the University of Melbourne. The following researchers kindly provided us with sequence alignments: Johannes Krause (*M. leprae*), Alexander Herbig (*M. leprae*), Zhemin Zhou (*Salmonella* Paratyphi A), Simon Harris (*B. pertussis*), Julian Parkhill (*B. pertussis*), Paul McAdam (*M. tuberculosis* Lineage 2, and *Staphylococcus aureus* CC398), Stephano Garasia (*K. pneumoniae*), Sarah Baines (*Staphylococcus aureus* ST239), Anne-Catrin Uhlemann (*Staphylococcus aureus* ST-USA300), Mark Davies (*Streptococcus pyogenes*), Mark Schultz (*A. baumannii*), Melissa Ward (*Staphylococcus aureus* CC398) and Iñaki Comas (*M. tuberculosis* New World mummies).

References

Baines, S. L., Holt, K. E., Schultz, M. B., Seemann, T., Howden, B. O., Jensen, S. O., van Hal, S. J., Coombs, G. W., Firth, N. & other authors (2015). Convergent adaptation in the dominant global hospital clone ST239 of methicillin-resistant *Staphylococcus aureus*. *MBio* **6**, e00080-15.

Bart, M. J., Harris, S. R., Advani, A., Arakawa, Y., Bottero, D., Bouchez, V., Cassiday, P. K., Chiang, C. S., Dalby, T. & other authors (2014). Global population structure and evolution of *Bordetella pertussis* and their relationship with vaccination. *MBio* **5**, e01074-14.

Biek, R., Pybus, O. G., Lloyd-Smith, J. O. & Didelot, X. (2015). Measurably evolving pathogens in the genomic era. *Trends Ecol Evol* **30**, 306–313.

Boni, M. F., Posada, D. & Feldman, M. W. (2007). An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics* **176**, 1035–1047.

Bos, K. I., Harkins, K. M., Herbig, A., Coscolla, M., Weber, N., Comas, I., Forrest, S. A., Bryant, J. M., Harris, S. R. & other authors (2014). Pre-Columbian mycobacterial genomes reveal seals as a source of new world human tuberculosis. *Nature* **514**, 494–497.

Bos, K. I., Herbig, A., Sahl, J., Waglechner, N., Fourment, M., Forrest, S. A., Klunk, J., Schuenemann, V. J., Poinar, D. & other authors (2016). Eighteenth century *Yersinia pestis* genomes reveal the long-term persistence of an historical plague focus. *Elife* **5**, e12994.

Bratcher, H. B., Corton, C., Jolley, K. A., Parkhill, J. & Maiden, M. C. (2014). A gene-by-gene population genomics platform: *de novo* assembly, annotation and genealogical analysis of 108 representative *Neisseria meningitidis* genomes. *BMC Genomics* **15**, 1138.

Bromham, L. (2009). Why do species vary in their rate of molecular evolution? *Biol Lett* **5**, 401–404.

Comas, I., Coscolla, M., Luo, T., Borrell, S., Holt, K. E., Kato-Maeda, M., Parkhill, J., Malla, B., Berg, S. & other authors (2013). Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nat Genet* **45**, 1176–1182.

Croucher, N. J., Harris, S. R., Fraser, C., Quail, M. A., Burton, J., van der Linden, M., McGee, L., von Gottberg, A., Song, J. H. & other authors (2011). Rapid pneumococcal evolution in response to clinical interventions. *Science* **331**, 430–434.

Croucher, N. J., Page, A. J., Connor, T. R., Delaney, A. J., Keane, J. A., Bentley, S. D., Parkhill, J. & Harris, S. R. (2014). Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* **43**, e15.

Cui, Y., Yu, C., Yan, Y., Li, D., Li, Y., Jombart, T., Weinert, L. A., Wang, Z., Guo, Z. & other authors (2013). Historical variations in mutation rate in an epidemic pathogen, Yersinia pestis. *Proc Natl Acad Sci U S A* **110**, 577–582.

Davies, M. R., Holden, M. T., Coupland, P., Chen, J. H., Venturini, C., Barnett, T. C., Zakour, N. L., Tse, H., Dougan, G. & other authors (2015). Emergence of scarlet fever *Streptococcus pyogenes* emm12 clones in Hong Kong is associated with toxin acquisition and multidrug resistance. *Nat Genet* **47**, 84–87.

Devault, A. M., Golding, G. B., Waglechner, N., Enk, J. M., Kuch, M., Tien, J. H., Shi, M., Fisman, D. N., Dhody, A. N. & other authors (2014). Second-pandemic strain of *Vibrio cholerae* from the Philadelphia cholera outbreak of 1849. *N Engl J Med* **370**, 334–340.

Didelot, X., Eyre, D. W., Cule, M., Ip, C. L., Ansari, M. A., Griffiths, D., Vaughan, A., O'Connor, L., Golubchik, T. & other authors (2012). Microevolutionary analysis of *Clostridium difficile* genomes to investigate transmission. *Genome Biol* **13**, R118.

Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* **29**, 1969–1973.

Duchêne, S., Holmes, E. C. & Ho, S. Y. W. (2014). Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. *Proc Biol Sci* **281**, 20140732.

Duchêne, S., Ho, S. & Holmes, E. C. (2015a). Declining transition/transversion ratios through time reveal limitations to the accuracy of nucleotide substitution models. *BMC Evol Biol* **15**, 36.

Duchêne, S., Duchêne, D., Holmes, E. C. & Ho, S. Y. W. (2015b). The performance of the date-randomization test in Phylogenetic analyses of time-structured virus data. *Mol Biol Evol* **32**, 1895–1906.

Eldholm, V., Monteserin, J., Rieux, A., Lopez, B., Sobkowiak, B., Ritacco, V. & Balloux, F. (2015). Four decades of transmission of a multidrug-resistant *Mycobacterium tuberculosis* outbreak strain. *Nat Commun* **6**, 7119.

Eyre-Walker, A. & Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nat Rev Genet* **8**, 610–618.

Firth, C., Kitchen, A., Shapiro, B., Suchard, M. A., Holmes, E. C. & Rambaut, A. (2010). Using time-structured data to estimate evolutionary rates of double-stranded DNA viruses. *Mol Biol Evol* **27**, 2038–2051.

Gaiarsa, S., Comandatore, F., Gaibani, P., Corbella, M., Dalla Valle, C., Epis, S., Scaltriti, E., Carretto, E., Farina, C. & other authors (2015).

Genomic epidemiology of *Klebsiella pneumoniae* in Italy and novel insights into the origin and global evolution of its resistance to carbapenem antibiotics. *Antimicrob Agents Chemother* **59**, 389–396.

Gibbs, M. J., Armstrong, J. S. & Gibbs, A. J. (2000). Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* **16**, 573–582.

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W. & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**, 307–321.

Hedge, J. & Wilson, D. J. (2014). Bacterial phylogenetic reconstruction from whole genomes is robust to recombination but demographic inference is not. *MBio* **5**, e02158–14.

Ho, S. Y. & Larson, G. (2006). Molecular clocks: when times are a-changin'. *Trends Genet* **22**, 79–83.

Ho, S. Y., Lanfear, R., Bromham, L., Phillips, M. J., Soubrier, J., Rodrigo, A. G. & Cooper, A. (2011). Time-dependent rates of molecular evolution. *Mol Ecol* **20**, 3087–3101.

Ho, S. Y. W., Duchêne, S., Molak, M. & Shapiro, B. (2015a). Time-dependent estimates of molecular evolutionary rates: evidence and causes. *Mol Ecol* **24**, 6007–6012.

Ho, S. Y. W., Duchêne, S. & Duchêne, D. (2015b). Simulating and detecting autocorrelation of molecular evolutionary rates among lineages. *Mol Ecol Resour* **15**, 688–696.

Holden, M. T., Hsu, L. Y., Kurt, K., Weinert, L. A., Mather, A. E., Harris, S. R., Strommenger, B., Layer, F., Witte, W. & other authors (2013). A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant *Staphylococcus aureus* pandemic. *Genome Res* **23**, 653–664.

Holt, K. E., Parkhill, J., Mazzoni, C. J., Roumagnac, P., Weill, F. X., Goodhead, I., Rance, R., Baker, S., Maskell, D. J. & other authors (2008). High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. *Nat Genet* **40**, 987–993.

Holt, K. E., Baker, S., Weill, F. X., Holmes, E. C., Kitchen, A., Yu, J., Sangal, V., Brown, D. J., Coia, J. E. & other authors (2012). *Shigella sonnei* genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nat Genet* **44**, 1056–1059.

Holt, K. E., Thieu Nga, T. V., Thanh, D. P., Vinh, H., Kim, D. W., Vu Tra, M. P., Campbell, J. I., Hoang, N. V., Vinh, N. T. & other authors (2013). Tracking the establishment of local endemic populations of an emergent enteric pathogen. *Proc Natl Acad Sci U S A* **110**, 17522–17527.

Holt, K., Kenyon, J. J., Hamidian, M., Schultz, M. B., Pickard, D. J., Dougan, G. & Hall, R. (2016). Five decades of genome evolution in the globally distributed, extensively antibiotic-resistant *Acinetobacter baumannii* global clone 1. *Microb Genomics* **2**, doi: 10.1099/mgen.0.000052.

Howden, B. P., Holt, K. E., Lam, M. M., Seemann, T., Ballard, S., Coombs, G. W., Tong, S. Y., Grayson, M. L., Johnson, P. D. & Stinear, T. P. (2013). Genomic insights to control the emergence of vancomycin-resistant enterococci. *MBio* **4**, e00412–00413.

Kay, G. L., Sergeant, M. J., Zhou, Z., Chan, J. Z.-M., Millard, A., Quick, J., Szikossy, I., Pap, I., Spigelman, M. & other authors (2015). Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. *Nat Commun* **6**, 6717.

Kennemann, L., Didelot, X., Aebischer, T., Kuhn, S., Drescher, B., Droege, M., Reinhardt, R., Correa, P., Meyer, T. F. & other authors (2011). *Helicobacter pylori* genome evolution during human infection. *Proc Natl Acad Sci U S A* **108**, 5033–5038.

Lapierre, M., Blin, C., Lambert, A., Achaz, G. & Rocha, E. P. (2016). The impact of selection, gene conversion, and biased sampling on the assessment of microbial demography. *Mol Biol Evol* **33**, 1711–1725.

Le Hello, S., Bekhit, A., Granier, S. A., Barua, H., Beutlich, J., Zajac, M., Münch, S., Sintchenko, V., Bouchrif, B. & other authors (2013). The global establishment of a highly-fluoroquinolone resistant *Salmonella enterica* serotype Kentucky ST198 strain. *Front Microbiol* **4**, 395.

Maixner, F., Krause-Kyora, B., Turaev, D., Herbig, A., Hoopmann, M. R., Hallows, J. L., Kusebauch, U., Vigl, E. E., Malfertheiner, P. & other authors (2016). The 5300-year-old *Helicobacter pylori* genome of the Iceman. *Science* **351**, 162–165.

Martin, D. & Rybicki, E. (2000). RDP: detection of recombination amongst aligned sequences. *Bioinformatics* **16**, 562–563.

Martin, D. P., Murrell, B., Golden, M., Khoosal, A. & Muhire, B. (2015). RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evol* **1**, vev003.

Marvig, R. L., Johansen, H. K., Molin, S. & Jelsbak, L. (2013). Genome analysis of a transmissible lineage of *pseudomonas aeruginosa* reveals pathoadaptive mutations and distinct evolutionary paths of hypermutators. *PLoS Genet* **9**, e1003741.

Merker, M., Blin, C., Mona, S., Duforet-Frebourg, N., Lecher, S., Willery, E., Blum, M. G., Rüsch-Gerdes, S., Mokrousov, I. & other authors (2015). Evolutionary history and global spread of the *Mycobacterium tuberculosis* Beijing lineage. *Nat Genet* **47**, 242–249.

Molak, M. & Ho, S. Y. W. (2015). Prolonged decay of molecular rate estimates for metazoan mitochondrial DNA. *PeerJ* **3**, e821.

Morelli, G., Song, Y., Mazzoni, C. J., Eppinger, M., Roumagnac, P., Wagner, D. M., Feldkamp, M., Kusecek, B., Vogler, A. J. & other authors (2010). *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. *Nat Genet* **42**, 1140–1143.

Murray, G. G., Wang, F., Harrison, E. M., Paterson, G. K., Mather, A. E., Harris, S. R., Holmes, M. A., Rambaut, A. & Welch, J. J. (2015). The effect of genetic structure on molecular dating and tests for temporal signal. *Methods Ecol Evol* **7**, 80–89.

Nelder, J. A. & Mead, R. (1965). A simplex method for function minimization. *Comput J* **7**, 308–313.

Njamkepo, E., Fawal, N., Tran-Dien, A., Hawkey, J., Strockbine, N., Jenkins, C., Talukder, K. A., Bercion, R., Kuleshov, K. & other authors (2016). Global phylogeography and evolutionary history of *Shigella dysenteriae* type 1. *Nat Microbiol* **1**, 16027.

O'Fallon, B. D. (2010). A method to correct for the effects of purifying selection on genealogical inference. *Mol Biol Evol* **27**, 2406–2416.

Padidam, M., Sawyer, S. & Fauquet, C. M. (1999). Possible emergence of new geminiviruses by frequent recombination. *Virology* **265**, 218–225.

Penny, D. (2005). Evolutionary biology: relativity for molecular clocks. *Nature* **436**, 183–184.

Pérez-Losada, M., Crandall, K. A., Zenilman, J. & Viscidi, R. P. (2007). Temporal trends in gonococcal population genetics in a high prevalence urban community. *Infect Genet Evol* **7**, 271–278.

Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol* **2**, vew007.

Rasmussen, S., Allentoft, M. E., Nielsen, K., Orlando, L., Sikora, M., Sjögren, K. G., Pedersen, A. G., Schubert, M., Van Dam, A. & other authors (2015). Early divergent strains of *Yersinia pestis* in Eurasia 5,000 years ago. *Cell* **163**, 571–582.

Rocha, E. P., Touchon, M. & Feil, E. J. (2006). Similar compositional biases are caused by very different mutational effects. *Genome Res* **16**, 1537–1547.

Salminen, M. O., Carr, J. K., Burke, D. S. & McCutchan, F. E. (1995). Identification of breakpoints in intergenotypic recombinants of

HIV type 1 by bootscanning. *AIDS Res Hum Retroviruses* **11**, 1423–1425.

Schuenemann, V. J., Singh, P., Mendum, T. A., Krause-Kyora, B., Jäger, G., Bos, K. I., Herbig, A., Economou, C., Benjak, A. & other authors (2013). Genome-wide comparison of medieval and modern *Mycobacterium leprae*. *Science* **341**, 179–183.

Schultz, M. B., Thanh, D. P., Do Hoan, N. T., Wick, R. R., Ingle, D. J., Hawkey, J., Edwards, D. J., Kenyon, J. J. & Lan, N. P. H. & other authors (2016). Repeated local emergence of carbapenem-resistant *Acinetobacter baumannii* in a single hospital ward. *Microbial Genomics* **2**, doi:10.1099/mgen.0.000050.

Smith, J. M. (1992). Analyzing the mosaic structure of genes. *J Mol Evol* **34**, 126–129.

Stinear, T. P., Holt, K. E., Chua, K., Stepnell, J., Tuck, K. L., Coombs, G., Harrison, P. F., Seemann, T. & Howden, B. P. (2014). Adaptive change inferred from genomic population analysis of the ST93 epidemic clone of community-associated methicillin-resistant *Staphylococcus aureus*. *Genome Biol Evol* **6**, 366–378.

Uhlemann, A. C., Dordel, J., Knox, J. R., Raven, K. E., Parkhill, J., Holden, M. T., Peacock, S. J. & Lowy, F. D. (2014). Molecular tracing of the emergence, diversification, and transmission of *S. aureus* sequence type 8 in a New York community. *Proc Natl Acad Sci U S A* **111**, 6738–6743.

Wagner, D. M., Klunk, J., Harbeck, M., Devault, A., Waglechner, N., Sahl, J. W., Enk, J., Birdsell, D. N., Kuch, M. & other authors (2014). *Yersinia pestis* and the plague of Justinian 541-543 AD: a genomic analysis. *Lancet Infect Dis* **14**, 319–326.

Ward, M. J., Gibbons, C. L., McAdam, P. R., van Bunnik, B. A., Girvan, E. K., Edwards, G. F., Fitzgerald, J. R. & Woolhouse, M. E. (2014). Time-scaled evolutionary analysis of the transmission and antibiotic resistance dynamics of *Staphylococcus aureus* donal complex 398. *Appl Environ Microbiol* **80**, 7275–7282.

Weller, C. & Wu, M. (2015). A generation-time effect on the rate of molecular evolution in bacteria. *Evolution* **69**, 643–652.

Xie, W., Lewis, P. O., Fan, Y., Kuo, L. & Chen, M. H. (2011). Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst Biol* **60**, 150–160.

Yahara, K., Didelot, X., Jolley, K. A., Kobayashi, I., Maiden, M. C., Sheppard, S. K. & Falush, D. (2016). The landscape of realized homologous recombination in pathogenic bacteria. *Mol Biol Evol* **33**, 456–471.

Zhou, Z., McCann, A., Litrup, E., Murphy, R., Cormican, M., Fanning, S., Brown, D., Guttman, D. S., Brisse, S. & Achtman, M. (2013). Neutral genomic microevolution of a recently emerged pathogen, *Salmonella enterica* serovar Agona. *PLoS Genet* **9**, e1003471.

Zhou, Z., McCann, A., Weill, F. X., Blin, C., Nair, S., Wain, J., Dougan, G. & Achtman, M. (2014). Transient Darwinian selection in *Salmonella enterica* serovar Paratyphi A during 450 years of global spread of enteric fever. *Proc Natl Acad Sci U S A* **111**, 12199–12204.

Data Bibliography

1. All the data sets used here are available online at: http://zenodo.org/record/45951#.Vr1M-JN95E4.