University of Sussex

Investigating Evolutionary Rate Variation in Bacteria

Bethany Gibson

Submitted for the degree of Master of Philosophy

November 2019

Declaration

I hereby declare that this thesis has not been and will not be, submitted in whole or in part to another University for the award of any other degree.

University of Sussex
Bethany Gibson, MPhil thesis
Investigating Evolutionary Rate Variation in Bacteria

Abstract

Rates of molecular evolution are known to vary between species and across all kingdoms of life. Here we explore variation in the rate at which bacteria accumulate mutations (accumulation rates) in their natural environments over short periods of time. We have compiled estimates of the accumulation rate for over 34 species of bacteria, the majority of which are pathogens evolving either within an individual host or during outbreaks. Across species we find that accumulation rates vary by over 3700-fold. We investigate whether accumulation rates are associated to a number potential correlates including genome size, GC content, measures of the natural selection and the time-frame over which the accumulation rates were estimated. After controlling for phylogenetic non-independence, we find that the accumulation rate is not significantly correlated to any factor. Furthermore, contrary to previous results we find that it is not impacted by the time-frame of which the estimate was made. We conclude that much of the rate variation is probably explained by variation in the generation time. We attempt to estimate doubling times of bacteria in the wild using a new method. We estimate the DT for five species of bacteria for which we have both an accumulation and a mutation rate estimate. We also infer the distribution of DTs across all bacteria from the distribution of the accumulation and mutation rates. Both analyses suggest that DTs for bacteria in the wild are substantially greater than those in the laboratory, that they vary by orders of magnitude between different species of bacteria and that a substantial fraction of bacteria double very slowly in the wild.

Acknowledgements

I would like to thank Professor Adam Eyre-Walker for his supervision and guidance, for sharing his time, immense knowledge, detailed feedback and continuous support. I thank Dr. Mark Paget for the supervision and helpful chats he provided. I would also like to thank Edward Feil, Daniel Wilson, and John Welch for collaboration, feedback and guidance.

I am deeply grateful to my friends and family who have given me amazing support and love throughout, especially my lab mate Jenny, my very good friends Lena, Chris, Dan and Ed, my parents Judi and Barry, and my brothers, David, Robin and Nathan. Finally, my greatest thanks to my partner Alex, who helped me through the good times and the bad and was forever there at my side.

Work in this thesis has been published as:

- Gibson, B., Wilson, DJ. Feil, E., & Eyre-Walker, A., 2018. The distribution of bacterial doubling times in the wild. *Proc. R. Soc. B,* 285: 20180789.
- Gibson, B. & Eyre-Walker, A., 2019. Investigating evolutionary rate variation in bacteria. *Journal of Molecular Evolution*, pp.1-10.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

1.AIMS

The goal of this of thesis is to investigate why molecular evolutionary rates vary across bacterial species. Evolutionary rates are known to vary across all kingdoms of life, including plants and animals. However, for bacteria, this topic remains relatively unexplored. This work aims to unravel the potential correlates of the accumulation rate in bacteria which will aid our understanding of bacterial evolution in general.

I first collect all available accumulation rate estimates from the literature and then to see if they correlate to several factors, including genome size, GC content, measures of natural selection and the time-frame over which the accumulation rates are measured. Secondly, I investigate whether another factor, generation time, can explain the variation in accumulation rates. To do this a new method is developed to estimate the generation time of bacteria in the wild. For this I need two sources of information: The accumulation rate and the mutation rate. Thus, further to collecting accumulation rates, I also collect mutation rates from the literature. I estimate doubling times for five species of bacteria and also the distribution of doubling times across all bacteria.

2. INTRODUCTION

Knowledge about the rates at which mutations arise and genomic change occurs is crucial to understanding how organisms evolve and adapt and how molecular evolution proceeds. Evolutionary rates are known to vary extensively across species in both prokaryotes and eukaryotes and this variation will in part be associated with species characteristics and biology. Disentangling the factors that influence evolutionary rates have been explored in many animal and plant systems (e.g. (Bromham 2002; Smith & Donoghue 2008; Welch *et al.* 2008; Lanfear *et al.* 2010), but not so much in bacteria (though see Rocha *et al.* 2006; Weller & Wu 2015; Duchêne et *al*. 2016). Here we investigate variation in the rate at which bacteria accumulate mutations through time in their natural environment over short time periods of a few months to a thousand years. We refer to these as accumulation rates to differentiate them from the mutation rate, the rate at which mutations occur, and the substitution rate, the rate at which mutations fix in a species. These rates of accumulation are commonly estimated using temporarily sampled data (Drummond *et al.* 2003), or concurrent samples from a population with a known date of origin (e.g. from fossil dates or co-speciation events). They vary by orders of magnitude from species such as *Mycobacterium leprae* with an accumulation rate of 8.6x10^{-9} (Schuenemann *et al*. 2013) to species such as *Campylobacter jejuni* with a rate of 3.23x10^{-5} (Wilson *et al*. 2009).

It remains unclear why the rate as which mutations accumulate varies so much between bacteria. The accumulation rate *per year* must ultimately depend upon the rate of mutation *per year* and the probability that a mutation reaches sufficient frequency in the population to be sampled. If some mutations are caused by DNA replication, as seems likely in most organisms, then the mutation rate *per year* is a function of the mutation rate *per generation* and the generation time. The probability that a mutation reaches a certain frequency in the population depends upon natural selection, biased gene conversion and the effective population size. We consider each of these explanations in turn.

It has previously been shown that the time-frame over which an accumulation rate is estimated can impact the estimate of evolutionary rate - they tend to be lower when measured over longer time-frames (Ho & Larson 2006; Ho *et al*. 2011; Duchene *et al*. 2014; Biek *et al*. 2015; Duchêne *et al.* 2016). This effect is usually attributed to the inefficiency of purifying selection to remove slightly deleterious mutations over shorter time periods or problems with reliably estimating rates when the sequences are saturated. This pattern is evident in bacteria (Rocha et al. 2006; Biek et al. 2015; Duchêne *et al.* 2016), however the evidence for the pattern is weak. In the most extensive analysis to date (Duchêne *et al.* 2016) the negative correlation between the accumulation rate and time-frame was a consequence of just two species which had been sampled over a long time-period. Furthermore, the authors removed datasets which showed no significant accumulation of mutations through time. This will have biased their analysis towards finding a negative correlation between the accumulation rate and sampling time-frame, because species with slow accumulation rates will be removed from the analysis if they are sampled over short-time frames, because they have not had enough time to accumulate significant numbers of substitutions.

Here we revisit the question of whether the accumulation rate is slow in species sampled over longer time-frames. We do this by comparing the rate of accumulation within species across different sampling times. We find little evidence for an association and consequently move on to explore other potential correlates of the accumulation rate. This includes 1) the mutation rate per generation, and 2) the effectiveness of selection. However, we find little evidence that these factors are responsible for the variation in the accumulation rate. This suggests that generation time might be a major factor.

Although, the generation time, or doubling rate, of bacteria has been measured in the lab for many species, relatively little is known about the DT of bacteria in their natural environment. For example,the bacterium *Escherichia coli* can divide every 20 minutes in the laboratory under aerobic, nutrient-rich conditions. But how often does it divide in its natural environment in the gut, under anaerobic conditions where it probably spends much of its time close to starvation? And what do we make of a bacterium, such as *Syntrophobacter fumaroxidans*, which only doubles in the lab every 140 hours (Harmsen *et al*. 1998). Does this reflect a slow doubling time in the wild, or our inability to provide the conditions under which it can replicate rapidly?

Estimating the generation time is difficult for most bacteria in their natural environment and very few estimates are available. The doubling time (DT) for intestinal bacteria has been estimated in several mammals by assaying the quantity of bacteria in the gut and faeces. Assuming no cell death Gibbons & Kapsimalis (1967) estimate the DT for all bacteria in the gut to be 48, 17 and 5.8 hours in hamster, guinea pig and mouse respectively. More recently Yang et al. (2008) have shown that the doubling time of *Pseudomonas aeruginosa* is correlated to cellular ribosomal content *in vitro* and have used this to estimate the DT *in vivo* in a cystic fibrosis patient to be between 1.9 and 2.4 hours.

We investigate what we can infer about the generation time in bacteria using a new method that uses two sources of information. First, the accumulation rate. If we assume that all mutations in the wild are neutral, an assumption that we show to be relatively unimportant for this method, in the discussion, then the accumulation rate is an estimate of the mutation rate per year, u_y. Second, we can estimate the rate of mutation per generation, u_g, in the lab using a mutation accumulation experiment and whole genome sequencing, or through fluctuation tests. If we assume that the mutation rate per generation is the same in the wild and in the lab, an assumption we discuss further below, then if we divide the accumulation rate per year in the wild by the mutation rate per generation in the lab, we can estimate the number of generations that the bacterium goes through in the wild and hence the doubling time (DT = 8760 x u_g / u_y , where 8760 is the number of hours per year).

In summary, we investigate why the rate of accumulation varies between bacterial species; we consider a number of explanations including the time-frame over which the estimates have been sampled, variation in the mutation rate and the efficiency of natural selection. We also attempt to estimate the generation time of bacterial in the wild, as a means to investigate whether variation in the generation time is a potential explanation for the variation in the rate of accumulation.

3. MATERIALS AND METHODS

3.1 Data collection

We compiled estimates of the accumulation rates from the literature (Appendix 1). For some species we obtained multiple estimates and in most analyses we use the average of these (Appendix 2). We also compiled estimates of the mutation rate from the literature and only used estimates that came from a mutation accumulation experiment with whole genome sequencing. If we had multiple estimates of the mutation rate, we summed the number of mutations across the mutation accumulation experiments and divided this by the product of the genome size and the number of generations that were assayed (Appendix 3). The genome size and GC content for each species is the average of all complete genomes on NCBI for each species. Nucleotide diversity estimates were calculated using orthologous sequence alignments for each species which were constructed using ODoSE ((Vos *et al.*, 2013),http://www.odose.nl) and in-house scripts written in Python (https://www.python.org) (Appendix 2). Lab Doubling times were taken from (Vieira-Silva & Rocha 2010) (Appendix 2).

We recalculated the accumulation rates in two cases in which the number of accumulated mutations had been divided by an incorrect number of years: *E. coli* (Reeves *et al*. 2011) and *Helicobacter pylori* (Kennemann *et al*. 2011). For *E. coli*, we reestimated the accumulation rate using BEAST by constructing sequences of the SNPs reported in the paper and the isolation dates. For, *Helicobacter pylori* we use two groups of strains in which strains were sampled from a patient at 0, 3 and 16-years; in both cases the 3-year and 16-year strains appear to form a clade to the exclusion of the 0-year strain because they share some common differences from the 0-year strain (Kennemann *et al*. 2011). We do not know when the 3-year and 16-year strains diverged, but assuming a molecular clock we can estimate the as follows: if the number of substitutions that have accumulated between the common ancestor of the 3-year and 16-year strain and each of the two strains are S_3 and S_{16} respectively then the rate of accumulation can be estimated as (S_{16}-S_3)/(13 years x genome size) (Figure 1.). Using the number of substitutions reported by (Kennemann *et al.* 2011) in their figure 1 we have estimated the accumulation rate to be 5 x 10^{-6} (for isolates NQ1707 and NQ4060) and 5.9 x 10^{-6} (for NQ1671 and NQ4191).

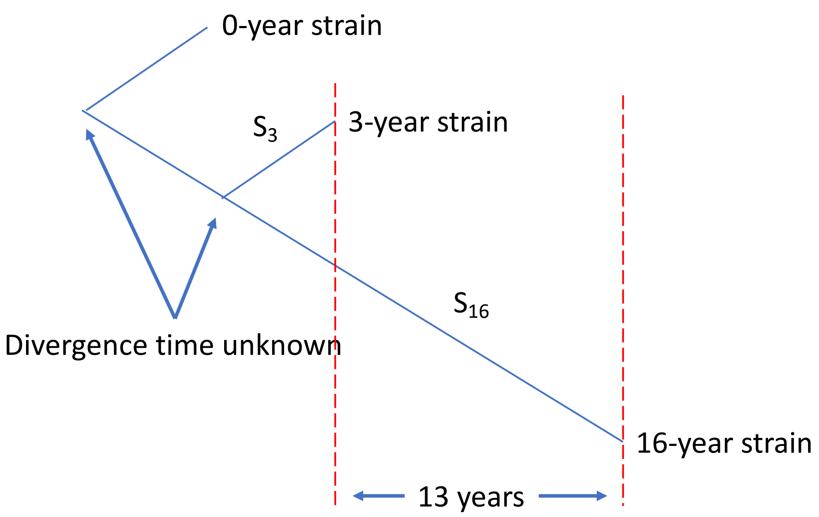


Figure 1. Estimating the accumulation rate for strains from Kennamann *et al.* 2011

We excluded some accumulation rate estimates for a variety of reasons. We only considered accumulation rates sampled over an historical timeframe of at most 1500 years. Most of our estimates of the accumulation rate are for all sites in the genome, so we excluded cases in which only the synonymous accumulation rate was given. We also excluded accumulation rates from hypermutable strains. Accumulation and mutation rate estimates used in the analysis are given in supplementary tables S1 and S2 respectively.

3.2 Testing for phylogenetic inertia

To estimate phylogenetic signal in the accumulation rates and all other traits we generated phylogenetic trees for the 34 species for which we have accumulation rate estimates (Appendix 5). 16S rRNA sequences were downloaded from the NCBI genome database (https://www.ncbi.nlm.nih.gov/genome/) and aligned using MUSCLE (Edgar 2004) performed in Geneious version 10.0.9 (http://www.geneious.com, Kearse *et al.*, 2012). From these alignments, maximum likelihood trees were constructed in RAxML (Stamatakis 2014) and integrated into the tests of Pagel (1999) and Blomberg et al. (2003) to the accumulation rates and all other traits implemented in the *phylosig* function in the R package *phytools v.0.6* (Revell 2012). Phylogenetic independent contrasts were carried out according to the method of Felsenstein (1985) using the *pic* function in *ape* v.4.1 (Paradis *et al.* 2004).

All statistical analyses were performed in R (https://cran.r-project.org).

3.3 Divergence as a function of time

The accumulation rate is expected to decrease as more divergent sequences are sampled because natural selection will remove deleterious genetic variation over time. To investigate this phenomenon quantitatively we used a transition matrix to explicitly calculate the distribution of allele frequencies *t* generations after a mutation was introduced into a haploid population. In the transition matrix the first column represents the population when the mutation is first introduced. If there are *N* strains (or chromosomes) in the population then there are *N*+1 rows, where the first row represents loss of the mutation and the *N*+1th row, fixation. The first column is therefore (0,1,0,0,0…0). To this column we apply selection and drift. If the fitness of the wildtype is 1 and the fitness of the mutant is 1-*s* then the frequency after selection is $f'(f,s) = (1-s)f/(1-sf)$ where *f* the frequency before selection. To calculate the effects of drift we use the binomial distribution. Hence the probability density of *x* copies of the mutation in generation *t* is

$$P(N,x,s,t) = \sum_{i=1}^{N-1} B(N,x,i,s)P(i,t-1) \qquad (1)$$

where *B*(*N,x,i,s*) is the binomial distribution taking into account the effects of selection

$$B(N,x,i,s) = \frac{N!}{x!(N-x)!}\left(f'(\tfrac{i}{N},s)\right)^{x}\left(1-f'(\tfrac{i}{N},s)\right)^{N-x} \qquad (2)$$

By applying equation 1 we can work out the probability density of a mutation introduced in the first generation in subsequent generations; i.e. we calculate *P*(*x,*2) for all *x* from 0 to *N*, and then *P*(*x*,3) for all *x* from 0 to *N*…etc). The *i*th column and *jth* row represent the probability of observing a mutation introduced as a single copy at generation 1, in *j* copies in the *i*th generation. The chance that a sequence sampled in *t* generations in the future is different to the ancestral can be calculated thus

$$D(N,s,t) = \sum_{v=1}^{t} \sum_{x=1}^{N} P(N,x,s,v)\frac{x}{N} \tag{3}$$

If we have two strains diverging from each other, then the overall divergence, assuming that mutations do not occur at the same site, which is reasonable for low levels of divergence, is twice this. We are interested in how selection affects the rate of accumulation and so we need to divide by the accumulation rate for neutral mutations, which is equivalent to dividing equation 3 by *t:*

$$A(N,s,t) = 2D(N,s,t)/2D(N,0,t) \tag{4}$$

In reality, not all deleterious mutations are subject to the same strength of selection so we sampled mutations from a gamma distribution; calculated *P*(*x,s,t*) for each and then averaged across mutations. We sampled 100 mutations for each set of parameters governing the distribution of fitness effects. $A(N,s,t)$ is expected to scale in *N* generations, something we have confirmed; i.e. $A(N,s,t) = A(zN,s,\frac{t}{z})$. We initially constructed a transition matrix with 100 strains to study the pattern from 0 to 4*N* generations, but then subsequently investigated the pattern in more depth within the first 0.1*N* generations by constructing a transition matrix with 1000 strains and the first 0.01*N* generations.

3.4 Estimating doubling times

We estimated the DT of individual species and the distribution across species using the formula DT = 8760 x u_g / u_y , where u_g is the mutation rate per generation as estimated from a mutation accumulation experiment, u_y is the mutation rate per year estimated from the accumulation rate, and 8760 is the number of hours per year. The estimate of the standard error associated with our estimate of the doubling time was obtained using the standard formula for the variance of a ratio: V(*x/y*) = (M(*x*)/M(*y*))^2(V(*x*)/M(*x*)^2+V(*y*)/M(*y*)^2) where M and V are the mean and variance of *x* and *y*. The variance for the accumulation rate was either the variance between multiple estimates of the accumulation rate if they were available, or the variance associated with the estimate if there was only a single estimate. The variance associated with the mutation rate was derived by assuming that the number of mutations was Poisson distributed.

To infer the distribution of DTs across bacteria we fit log-normal distributions to the accumulation and mutation rate data by taking the $\log_e$ of the values and then fitting a normal distribution by maximum likelihood using the *FindDistributionParameters* in *Mathematica*. Normal Q-Q plots for the accumulation and mutation rate data were produced using the qqnorm function in R version 1.0.143. In fitting these distributions, we have not taken into account the sampling error associated with the accumulation and mutation rate estimates. However, these sampling errors are small compared to the variance between species: for the accumulation rates the variance between species is 3.9 x 10^{-11} and the average error variance is an order of magnitude smaller at 3.6 x 10^{-12}; for the mutation rate data, the variance between species is 7.5 x 10^{-18} and the average variance associated with sampling is more than two orders of magnitude smaller at 1.8 x 10^{-20}. Note that we cannot perform these comparisons of variances on a log-scale because we do not have variance estimates for the log accumulation and mutation rates.

4. RESULTS

4.1 Across species

We compiled estimates of the accumulation rate for 34 species of bacteria. These vary by over 3700-fold (Figure. 2.), but the majority of species accumulate mutations at rates of between 1x10^{-6} and 2x10^{-6} per site per year. In the sections below, we investigate what might cause this variation by looking for variables which correlate to the accumulation rate. Because the accumulation rate varies over orders of magnitude, all analyses were performed on the log of the accumulation rate. In such an analysis it can be important to correct for phylogenetic non-independence if there is a phylogenetic inertia. To investigate this we tested for phylogenetic inertia by inferring the phylogeny of our species using the 16S rRNA and then using the tests of Pagel (1999) and Blomberg, Garland and Ives (2003). We find that the accumulation rates show phylogenetic inertia using Pagel's λ but not Bloomberg *et al.*'s K , and some of our other variables also show inertia including genome size and GC content, but not all (Table 1).

Trait	**λ**	**p value**	**K**	**p value**
Mutation Rate	0.88	0.026	0.5	0.009
Accumulation Rate	0.68	0.001	0.0005	0.37
Genome size	1	<0.001	0.38	0.001
GC content	1	<0.001	0.79	0.001
$\pi N/\pi S$	0.000062	0.99	0.0077	0.108
Lab DT	0.8	0.003	0.08	0.279

Table 1. Tests of phylogenetic signal. Pagel's λ (Pagel 1999) and Blomberg *et al.*'s K (Blomberg *et al.* 2003).

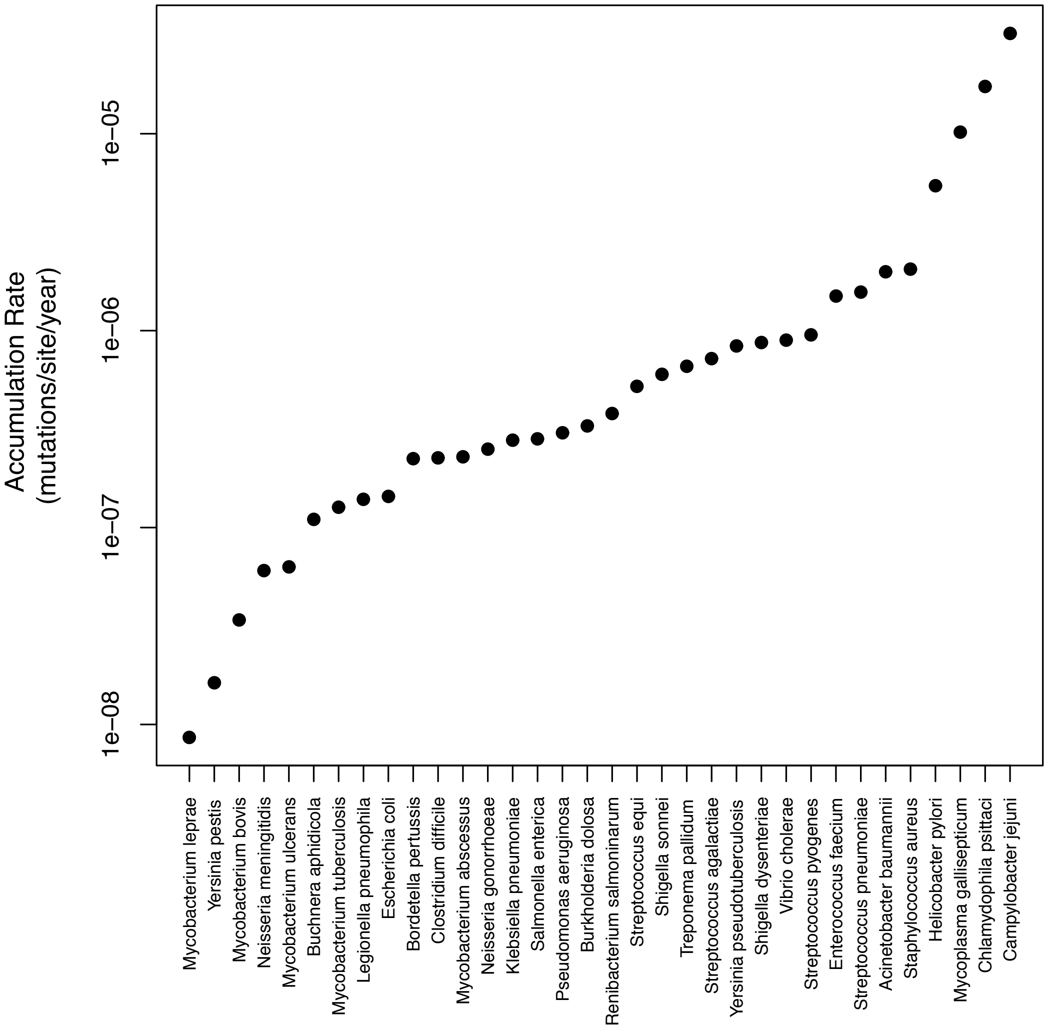


Figure 2. Distribution of accumulation rate estimates for 34 species of bacteria.

4.2 Sampling time

The time-interval over which evolutionary rates are measured is thought to impact rate estimates so that they become slower when measured over longer time-frames (Ho *et al*. 2011; Biek *et al.* 2015; Duchêne et al. 2016). This is as we might expect if a substantial fraction of mutations are mildly deleterious, since they would appear over a short time-scale, but ultimately be removed by natural selection. Evidence for this effect comes from observation that the relative rate at which non-synonymous and synonymous mutations accumulate in bacterial genomes declines as a function of time (Rocha *et al*. 2006; Balbi & Feil 2007).

We test whether the accumulation rate estimates scale negatively with sampling time, defined here as either the time-interval over which isolates were temporally sampled or the divergence time separating concurrent sequences. Sampling time varies from 1 year to just over 1500 years. We find a highly significant negative relationship between accumulation rate and sampling time (Figure 3.) (r = -0.38, *p* = 0.0016) across all species for all studies, but this appears to be largely contributed by four points associated with two species, *Yersinia pestis* and *Mycobacterium leprae*. It is not clear whether *Y. pestis* and *M. leprae* have low rates because this is a feature of their evolution, irrespective of the time frame over which they were sampled, or because they have been sampled over long time frames. For several species there are multiple estimates of the accumulation rate. If we control for any species effects by considering the correlation between the accumulation rate and the sampling time-frame within these 12 species using ANCOVA, we find no correlation (slope = 0.022, *p* = 0.79) (Figure. 3). Furthermore, we find no relationship between the relative rates at which non-synonymous and synonymous mutations accumulate and the time-frame over which the accumulation rate estimate was made (r = 0.2, p = 0.53), although for most datasets the accumulation rate was not calculated for the two types of site separately. In conclusion, we do not find strong evidence for a sampling time effect.

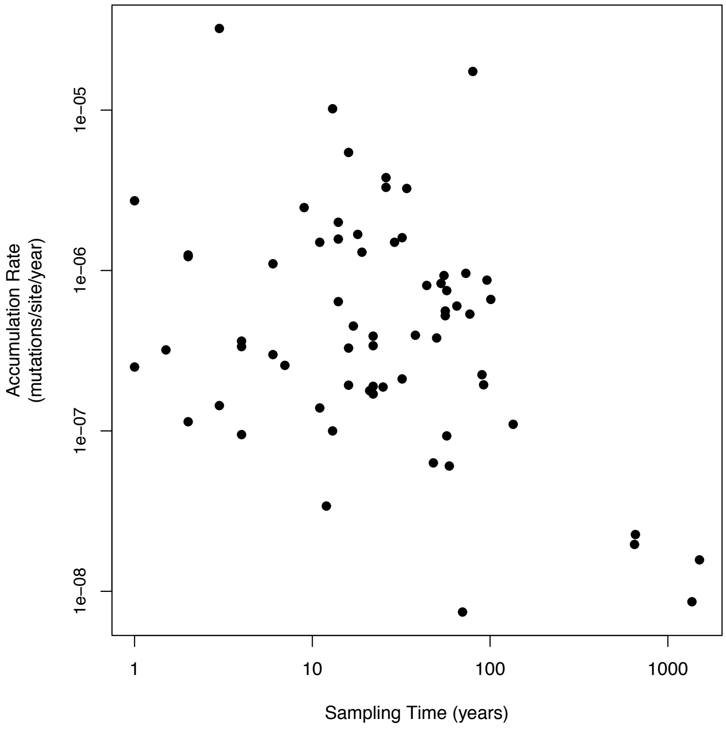


Figure 3. The accumulation rate vs sampling time.

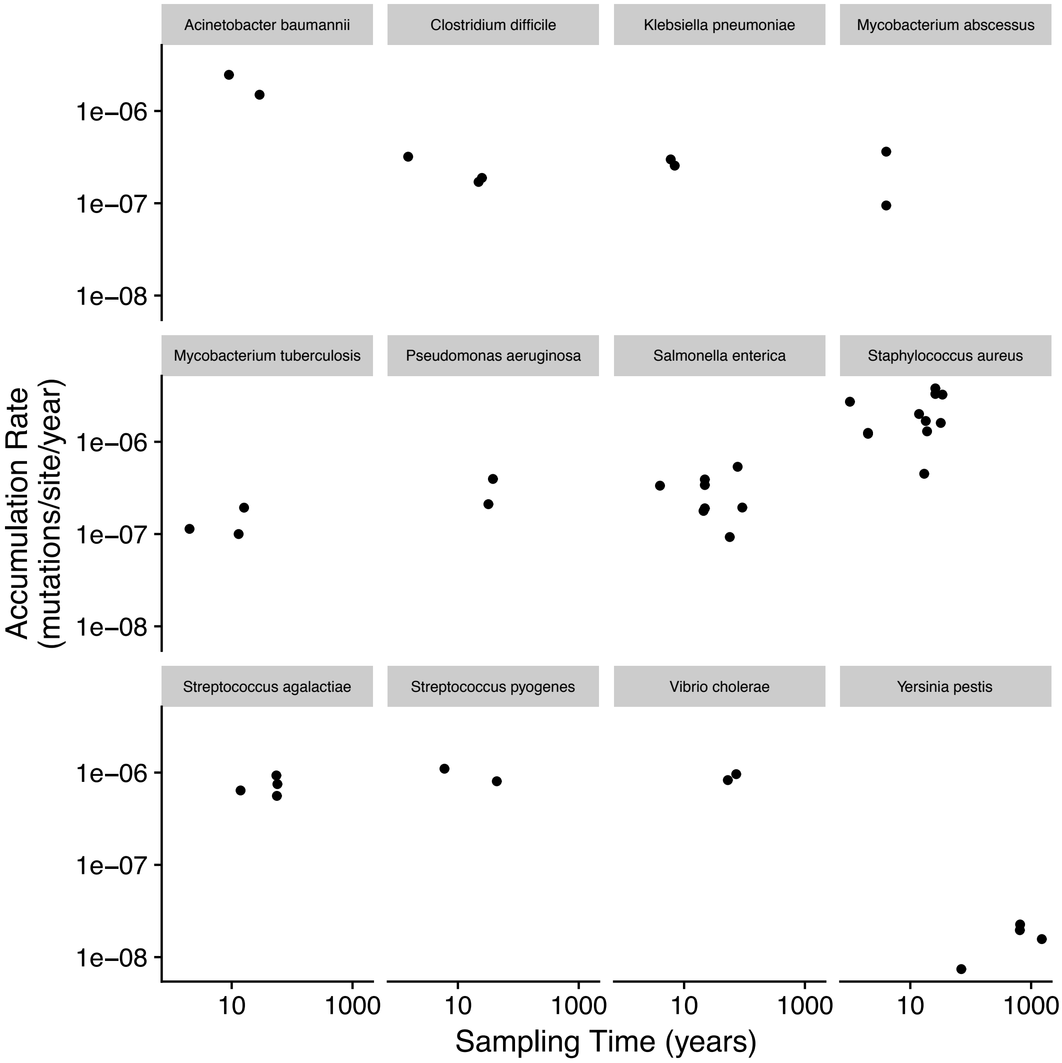


Figure 4. The accumulation rate vs sampling time split for the 12 species for which we have multiple estimates.

The absence of a relationship between the accumulation rate and sampling time might seem surprising given that there is ample evidence that slightly deleterious mutations segregate in bacterial populations; for example, (Hughes 2005) showed that non-synonymous polymorphisms segregate at lower frequencies than synonymous polymorphisms in most species of bacteria. So, we would expect the rate of accumulation to decline as time progresses. To investigate this further, we derived the expected relationship between the accumulation rate and time using population genetic theory (see Materials and Methods). We assume all mutations are drawn from a distribution of fitness effects (DFE), modelled as a gamma distribution, in which all mutations are either effectively neutral, or deleterious. We find, as expected, that the rate of accumulation declines. However, it is evident that it will be difficult to detect differences in accumulation rate unless accumulation rates are sampled over a very short time frame (<0.1*N* generations, where *N* is the population size) and a much longer time frame (Figure 5). This is because within a restricted time frame there is very little difference in accumulation rate.

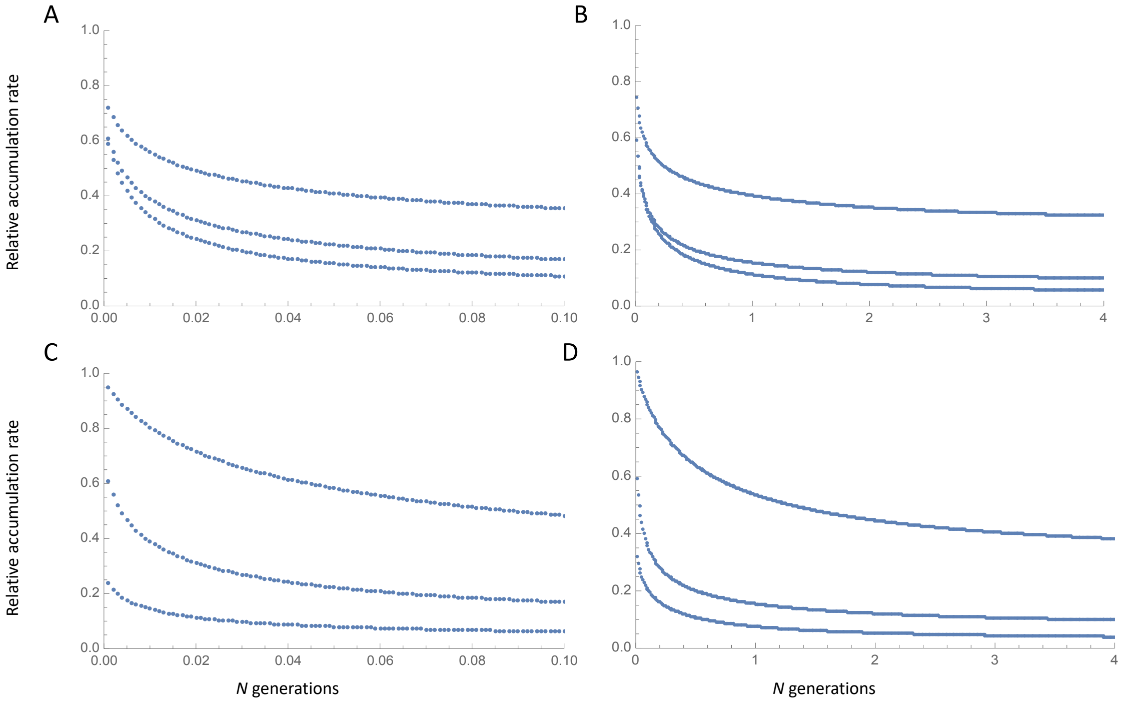


Figure 5. The expected relationship between the accumulation rate at selected sites relative to neutral sites and sampling time. In panels A and B the shape parameter of the gamma distribution is varied 0.25 (top line), 0.50 (middle) and 0.75 (bottom); in panels C and D the mean strength of selection, multiplied by *N*, is varied from 10 (top), 100 (middle) and 1000 (bottom). Panels A and C show the relative accumulation rate over the first 0.1*N* generations, panels B and D over the first 4*N* generations.

4.3 Mutation rate

The rate at which bacteria accumulate mutations through time will be in part be determined by the rate at which mutations occur per unit time. If some mutations are caused by DNA replication then the mutation rate *per year* will depend upon the mutation rate *per generation* and the generation time. We test each of these components in turn.

Unfortunately, it is difficult to directly test for a relationship between the accumulation rate and the mutation rate per generation because only five species in our dataset have estimates of both these rates. The correlation between the accumulation rate and mutation rate per generation is 0.07 (*p*=0.9), but with such little information it is difficult to determine whether a correlation exists. However, it is potentially possible to test the relationship between the accumulation rate and the mutation rate per generation indirectly because some genomic traits correlate to the mutation rate per generation. For instance, genome size is inversely correlated to the mutation rate/site/generation (Drake 1991; Lynch 2010; Lynch, Matthew S. Ackerman, *et al*. 2016) . We find a negative relationship between the mutation rate and genome size (r= -0.68, p= <0.001), although this is mostly driven by *Mesoplasma florum* (Appendix 6.) and the correlation is weaker when we remove *M.florum* (r = -0.39, *p*= 0.053).

A negative correlation between genome size and the accumulation rate has been previously observed for a range of viruses and bacteria (Lynch 2010; Biek *et al*. 2015) and we also find a strong negative correlation between the accumulation rate and genome size (Figure 6a) (r = -0.43 , *p*=0.01) which becomes stronger when the obvious outlier *B. aphidicola* is excluded (r = -0.57, *p* = <0.001). The relationship is also negative, but loses significance, if we control for phylogeny using phylogenetic independent-contrasts (PICs) after excluding low variance comparisons and *B.aphidicola* (r=-0.27, p=0.23) (Figure. 6b). 10 comparisons were considered low variance as their standard deviations were <0.21.

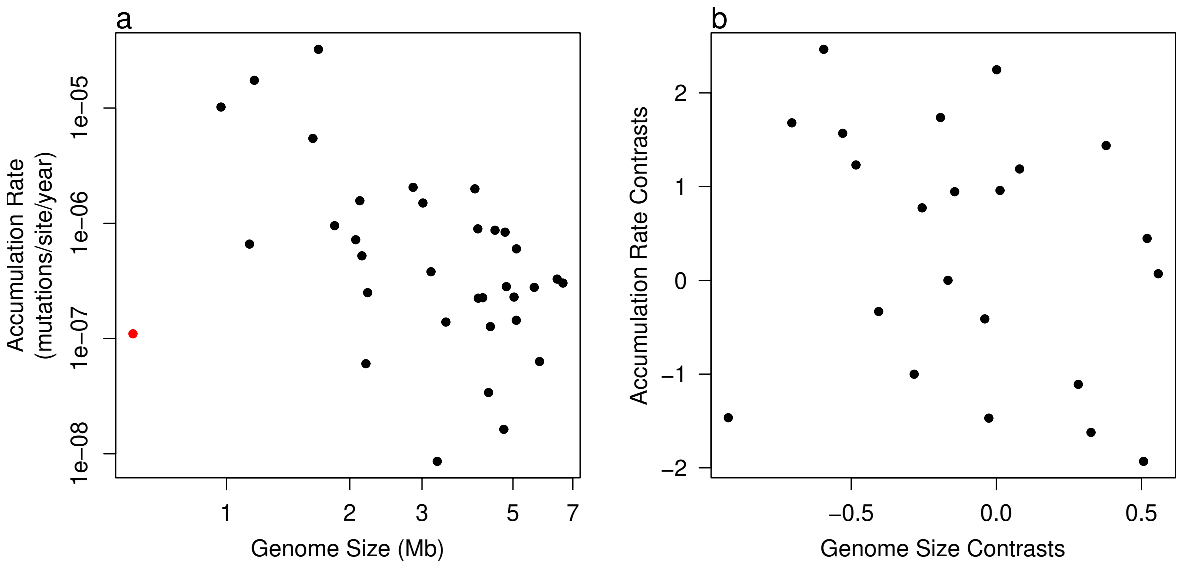


Figure 6. **a)** The accumulation rate vs genome size. The outlier, *Buchnera aphidicola,* is highlighted in red **b)** phylogenetic independent contrasts for the accumulation rate vs genome size.

Genomic base composition may also correlate to the mutation rate per generation. GC content is known to vary greatly across bacterial species from less than 20% to over 70%. The origins of this variation remain unresolved. There is evidence that it is not solely a consequence of mutation bias (Hildebrand *et al.* 2010; Hershberg & Petrov 2010) and that biased gene conversion may be a factor (Lassalle *et al.* 2015). Given that the pattern of mutation is generally AT-biased in bacteria (Hershberg & Petrov 2010) (though see Long *et al.*, 2015; Sun *et al.*, 2017) variation in GC content due to selection or biased gene conversion can potentially generate variation in the mutation rate by shifting the GC-content away from its equilibrium value (Krasovec *et al.* 2017). This effect may explain why *Mesoplasma florum*'s mutation rate is so high because although it has very low genomic GC content, the equilibrium GC content is predicted to be substantially lower (Krasovec *et al.* 2017). This will lead to positive correlation between the accumulation rate and GC content. The mutation rate may also be negatively correlated to GC-content due to variation in effective population size; a low effective population size may lead to lower GC content but a higher mutation rate because selection on mutation rate modifiers is relaxed and repair genes are lost. Surprisingly we find a positive association between an inverse measure of $N_e(\pi_N/\pi_S)$

and GC content (r=0.473, p=0.0094), although this is lost when we account for phylogenetic non-independence (r=0.32, p=0.168).

We observe a negative correlation between GC content and the mutation rate (r=-0.59, *p* = 0.0016) (Appendix 7.), and we also find a strong negative correlation between the accumulation rate and the GC-content (r = -0.53 *p*= 0.001; Fig. 6a). Again, *B. aphidicola* is a conspicuous outlier and if removed the correlation is stronger (r = -0.613, *p*=<0.001). This negative relationship is maintained and is almost significant for phylogenetic independent-contrasts (-0.390, *p*=0.072) after exclusion of *B.aphidicola* and low variance points (Figure. 7b).

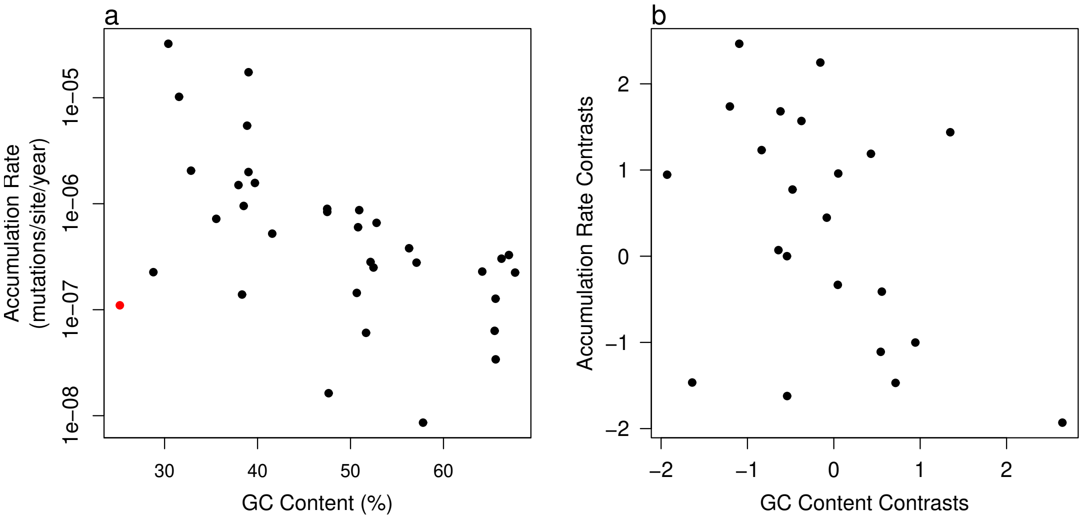


Figure 7. **a)** The accumulation rate vs GC content. The outlier, *Buchnera aphidicola,* is highlighted in red. **b)** phylogenetic independent contrasts for the accumulation rate vs GC content.

We have detected moderately significant correlation between the accumulation rate and genome size and GC-content. These two variables are correlated to each other but a multiple regression of accumulation rate versus both yields marginally significant results for GC content (*p*=0.037) but not significant for genome size (*p*=0.45) and neither come out significant when we control for phylogeny; it is therefore not possible for us to clearly resolve which might be the true correlate. Both could conceivably be linked to the mutation rate per generation. Under the drift-limit hypothesis the mutation rate is expected to be negatively correlated to genome size, because larger genomes have potentially more deleterious mutations and this leads to more effective selection on the mutation rate (Lynch 2010; Lynch 2017). GC-content could be related to the mutation rate either through its correlation to genome size, a correlation for which there is no clear explanation, or because GC-content is a crude measure of how far a genome is from its equilibrium GC-content; if the mutation pattern is AT-biased then increasing GC-content increases the mutation rate (Krasovec 2017).

4.4 Effectiveness of selection

Selection and biased gene conversion will affect the probability that a mutation spreads to fixation in a population. Accumulation rates are estimated by excluding sites which are inferred to have been recombined and hence biased gene conversion is unlikely to explain the variation. In contrast, purifying selection will act to reduce the number of deleterious mutations surviving in populations, leading to a reduction in the accumulation rate. How effective selection is at exerting its effects depends on the power of random genetic drift, i.e. the effective population size. We can potentially measure the effectiveness of selection by considering the ratio of the nucleotide diversity at non-synonymous and synonymous sites (πN/πS); populations with more efficient selection should have lower values of πN/πS. We consider the efficiency of selection using two sources of data; the ratio of the number of non-synonynous to synonymous polymorphisms, pN/pS, for the strains used to estimate the accumulation rate and πN/πS in the species as a whole. We find no correlation between pN/pS in the strains to estimate the accumulation rate (r=0.07, p =0.84) but we have only nine data-points. We find an almost significant correlation for the species wide πN/πS and the accumulation rate (r= -0.35, p=0.062) but none if we control for phylogenetic inertia. (r = -0.1, *p*=0.65).

4.5 Lifestyle

We examined whether there are differences in the accumulation rate for bacteria with different lifestyles. Most of our species are pathogens and among these we divided them into obligate pathogens and opportunistic pathogens. We find that the accumulation rates do not differ significantly between these two groups (t-test, *p*=0.488). We further carried out an analysis controlling for phylogenetic non-independence by comparing sister pairs of species. We find no evidence that they are significantly different (paired sample t-test, *p*=0.947). Thus, lifestyle does not seem to have any clear impact on the accumulation rate.

4.6 All factors

We further carried out a multivariate analysis where we included all our variables into a multiple regression (apart from our estimates of DTs in the wild). When we consider the raw values, only genome size comes out as significant (*p*= 0.0153) and when we consider the phylogenetic independent contrasts lab doubling times and πN/πS come out as marginally significant with similar effect sizes (Standardized regression coefficient = -0.095, *p*=0.080 and 1.01, *p* = 0.063 respectively); this suggests that accumulation rates may be higher in species with short lab DTs and smaller N_e.

4.7 Generation time

It is likely that the accumulation rate should correlate negatively with generation time (or doubling time) because species with shorter generation times will accumulate more DNA replication errors per unit time. Eukaryotes appear to display this generation time-effect (Bromham 2002; Smith & Donoghue 2008; Welch *et al*. 2008; Lanfear *et al*. 2010) and this is also evident in bacteria (Weller and Wu 2015) although see (Maughan 2007). Furthermore, the accumulation rate may also increase in populations that are rapidly expanding, for instance during epidemic disease, because of a reduction in generation time (Cui *et al.* 2013).

However, we find no relationship between the accumulation rate and the doubling time, as measured in the lab (r=-0.483, *p* =.0.60 for raw values and r = -0.298, *p* = 0.21 for phylogenetic independent contrasts). Other genomic features also correlate to lab doubling times (Vieira-Silva and Rocha 2010) but we find no correlation between the accumulation rate and 16s gene copy number (r= 0.044, *p*= 0.802 for raw values and r= 0.126, *p*= 0.565 for phylogenetic independent contrasts) and tRNA abundance (r= -0.085,*p*= 0.63 for raw values and r= 0.156, *p* = 0.47 for phylogenetic independent contrasts). This may be because lab doubling times do not reflect what occurs in the wild but they might relate to some aspect of bacteria life history.

Unfortunately, there are very few estimates of the DT of bacteria in their natural environment. However, we can use an indirect method to potentially estimate the DT. If we assume that the mutation rate per generation is the same in the lab and in the wild, and that all mutations are neutral, then dividing the accumulation rate per year by the mutation rate per generation in the lab, yields an estimate of the number of generations per year, and hence the DT.

The accumulation rate in the wild and the mutation rate in the lab have been estimated for 34 and 26 bacterial species respectively (Tables A2, A3); we only consider mutation rate estimates from mutation accumulation experiments, since estimates from fluctuation tests are subject to substantial sampling error and unknown bias, and we exclude estimates from hypermutable strains. For five species, *Escherichia coli, Pseudomonas aeruginosa, Salmonella enterica, Staphylococcus aureus and Vibrio cholerae,* we have both an accumulation and a mutation rate estimate and hence can estimate the DT. Amongst these five species we find our DT estimates vary from 1.1 hours in *V. cholerae* to 25 hours in *Salmonella enterica* (Table 2). In all cases the estimated DT in the wild is greater than that of the bacterium in the lab. For example, *E. coli* can double every 20 minutes in the lab but we estimate that it only doubles every 15 hours in the wild.

In theory, it might be possible to estimate the DT in those bacteria for which we have either an accumulation or mutation rate estimate, but not both, by finding factors that correlate with either rate and using those factors to predict the rates. Unfortunately, we have been unable to find any factor that correlates sufficiently well to be usefully predictive. As mentioned it has been suggested that the mutation rate is correlated to genome size in microbes (Drake 1991) but, the current evidence for this correlation is

Species	Accumulation rate per site per year	Mutation rate per site per generation	DT (hr) (SE)	Lab DT (hr)	Ratio	AR Ref	MR Ref
Escherichia coli	1.44 x 10^{-7}	2.54 x 10^{-10}	15 (7.7)	0.33	45	1	6
Pseudomonas aeruginosa	3.03 x 10^{-7}	7.92 x 10^{-11}	2.3 (0.77)	0.5	4.6	2	7
Salmonella enterica	2.50 x 10^{-7}	7.00 x 10^{-10}	25 (7.9)	0.5	50	3	8
Staphylococcus aureus	2.05 x 10^{-6}	4.38 x 10^{-10}	1.87 (0.98)	0.4	4.7	4	9
Vibrio cholerae	8.30 x 10^{-7}	1.07 x 10^{-10}	1.1 (0.26)	0.66	1.7	5	10

Table 2. Doubling time estimates (hours) for those species for which we have both an estimate of the accumulation and mutation rate. Accumulation rate (AR) references – 1) (Reeves *et al*. 2011); 2) (Markussen & Marvig 2014; Marvig *et al*. 2013); 3) (Duchêne *et al.* 2016; Zhou *et al.* 2014; Okoro *et al*. 2012; Hawkey *et al*. 2013; Mather 2013); 4) (Stinear *et al.* 2014; Baines *et al.* 2015; Ward *et al*. 2014; Holden *et al.* 2013; Uhlemann *et al*. 2014; Alam *et al.* 2015; Smyth *et al*. 2010; Harris *et al.* 2010; Gray *et al*. 2011; Nübel *et al.* 2010; Young *et al*. 2012); 5) (Mutreja *et al*. 2011; Duchêne *et al*. 2016). Mutation rate (MR) references – 6) (Lee *et al*. 2012); 7) (Dettman *et al.* 2016); 8) (Lind & Andersson 2008); 9) (Long *et al.* 2018); 10) (Sung *et al*. 2012).

very weak, and depends upon the estimate from *Mesoplasma florum* (r =-0.68, *p* < 0.001 with *M. florum* and r = -0.39, *p*= 0.053 without *M. florum*) (Appendix 6.) (Lynch, Matthew S Ackerman, *et al.* 2016). However, we can use the accumulation and mutation rate estimates to estimate the distribution of DTs across bacteria if we assume that there is no phylogenetic non-independence in the mutation and accumulation data, an assumption we address below. We can estimate the distribution of DTs by fitting distributions to the accumulation and mutation rate data, using maximum likelihood, and then dividing one distribution by the other. We assume that both variables are log-normally distributed, an assumption which is supported by Q-Q plots with the exception of the mutation rate per generation in *Mesoplasma florum*, which is a clear outlier (Figure 8.). We repeated all our analyses with and without *M. florum.*

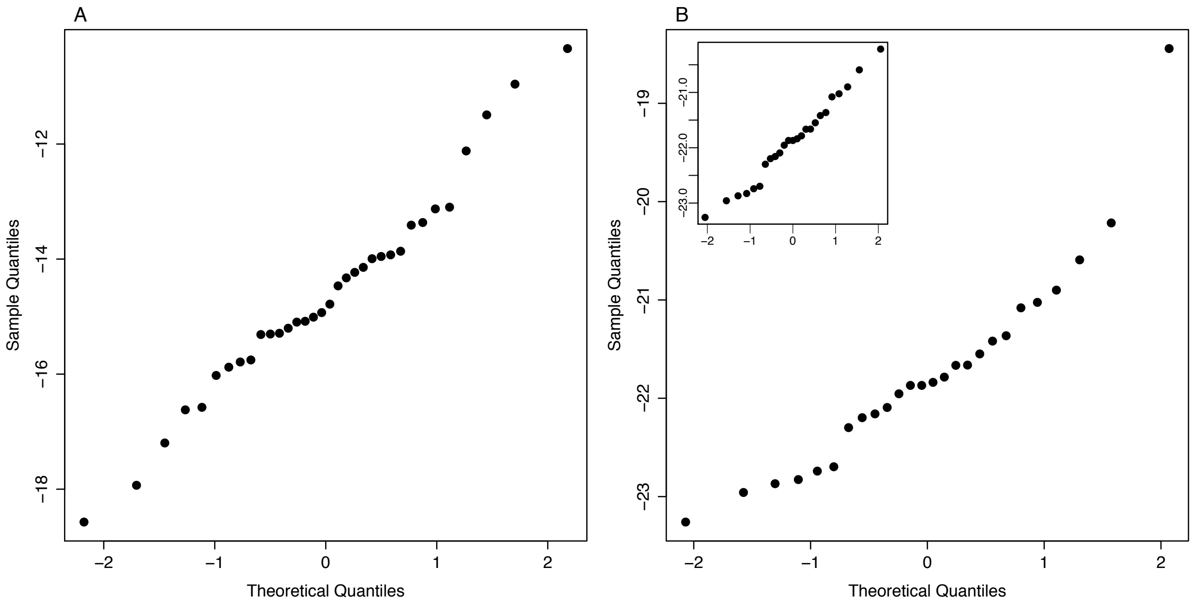


Figure 8. Normal Q-Q plots for the log of (A) accumulation and (B) mutation rate data. The main plot in B includes all twelve mutation rate estimates and the insert excludes *Mesoplasma florum* estimate.

If the accumulation and mutation rate data are log-normally distributed then the distribution of DT is also log-normally distributed with a mean of $\log_e(8760) + m_g - m_y$ and a variance of $v_g + v_y - 2\mathrm{Cov}(g,y)$, where 8760 is the number of hours per year and m_g, m_y, v_g and v_y are the mean and variance of the lognormal distributions fitted to the mutation (subscript *g)* and accumulation (subscript *y*) rates. *Cov*(*g,y*) is the covariance between the accumulation and mutation rates. We might expect that species with higher mutation rates also have higher accumulation rates, because the accumulation rate is expected to depend on the mutation rate, but the correlation between the two will depend upon how variable the DT and other factors, such as the strength of selection, are between bacteria. The observed correlation between the log accumulation rate and log mutation rate is 0.077, but there are only five data points, so the 95% confidence intervals on this estimate encompass almost all possible values (-0.86 to 0.89). We explore different levels of the correlation between the accumulation and mutation rates; it should be noted that *Cov*(*g,y*) can be expressed as $\mathrm{Sqrt}(v_g\, v_y)$ *Corr*(*g,y*) where *Corr*(*g,y*) is the correlation between the two variables.

The distribution of DTs in the wild inferred using our method is shown in Figure 8. We infer the median doubling time to be 7.04 hours, but there is considerable spread around this even when the accumulation and mutation rates are strongly correlated (Figure 8A); as the correlation increases so the variance in DTs decreases, but the median remains unaffected. The analysis suggests that most bacteria have DTs of between 1 and 100 hours but there are substantial numbers with DTs beyond these limits. For example, even if we assume that the correlation between the accumulation and mutation rate is 0.5 we infer that 10% of bacteria have a DT of faster than one hour in the wild and 4.2% have a DT slower than 100 hours in the wild. If we remove the *Mesoplasma florum* mutation rate estimate from the analysis the median doubling is slightly lower at 6.16 hours, but there is almost as much variation as when this bacterium is included; at a correlation is 0.5 we infer that 12% of bacteria have a DT faster than one hour in the wild and 3.5% have a DT slower than 100 hours.

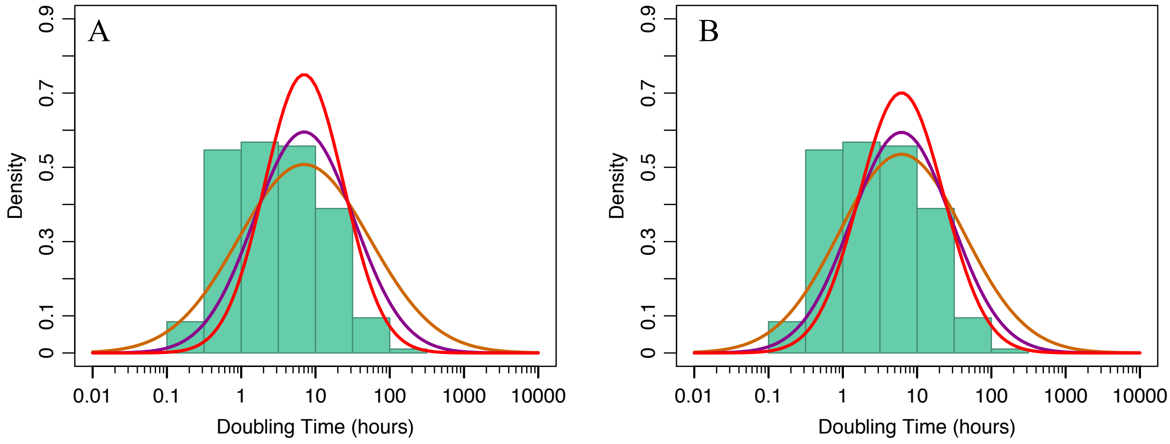


Figure 9. The distribution of DTs amongst bacteria inferred assuming different levels of correlation between the accumulation and mutation rates - orange r = 0, purple r = 0.5 and red r = 0.75. We also show the distribution of lab DTs (green histogram) from a compilation of over 200 species made by Vieira-Silva and Rocha (2010). In panel A we include all mutation rate estimates and in panel B we exclude the mutation rate estimate for *Mesoplasma florum*.

To investigate how robust these conclusions are to statistical sampling, we bootstrapped the accumulation and mutation rate estimates, refit the log-normal distributions and reinferred the distribution of DT. The 95% confidence intervals for the median are quite broad at 3.4 to 14.2 hours (3.1 to 11.3 hours if we exclude *M. florum*). However, all bootstrapped distributions show substantial variation in the DT with a substantial fraction of bacteria with long DTs and also some with very short DTs (Figure 10).

Here, we have assumed that there is no phylogenetic inertia within the accumulation and mutation rate estimates. As stated above to test whether this is the case we constructed a phylogenetic tree using 16S rRNA sequences and applied the tests of Pagel (1999) (Pagel 1999) and Blomberg *et al.* (2003) (Blomberg *et al.* 2003). Both the accumulation and mutation rate data show some evidence of phylogenetic signal. For the accumulation data, Pagel's λ = 0.68 (p = 0.001) and Blomberg *et al*.'s K = 0.0005 (p = 0.35); and for the mutation rate data Pagel's λ = 0.88 (p = 0.026) and Blomberg *et al.'s* K = 0.5 (p = 0.009). We also find some evidence that the data depart from a Brownian motion model using Pagel's test (i.e. λ is significantly less than one) for the accumulation data (p<0.001) but not the mutation rate data (p = 0.094); i.e. the accumulation rates are more different than we would expect from their phylogeny and a Brownian motion model. A visual inspection of the data suggests that the phylogenetic signal is largely contributed by species that are closely related, rather than deeper phylogenetic levels (Figure 11A, B) and species for which we have accumulation and mutation rate estimates are interspersed with one another on the phylogenetic tree (Figure 11C). It therefore seems unlikely that phylogenetic inertia will influence our results.

It is of interest to compare the distribution of DTs in the wild to the distribution of lab DTs (Figure 9). The distributions are different in two respects. First, the median lab DT of 3 hours is less than half the median wild DT of 7.04 hours (6.16 hours without *M. florum*); the two are significantly different (p = 0.012 with *M. florum* and p = 0.016 without *M.florum*, inferred by bootstrapping each dataset and recalculating the medians). Second, many more bacteria are inferred to have long DTs in the wild than in the lab.

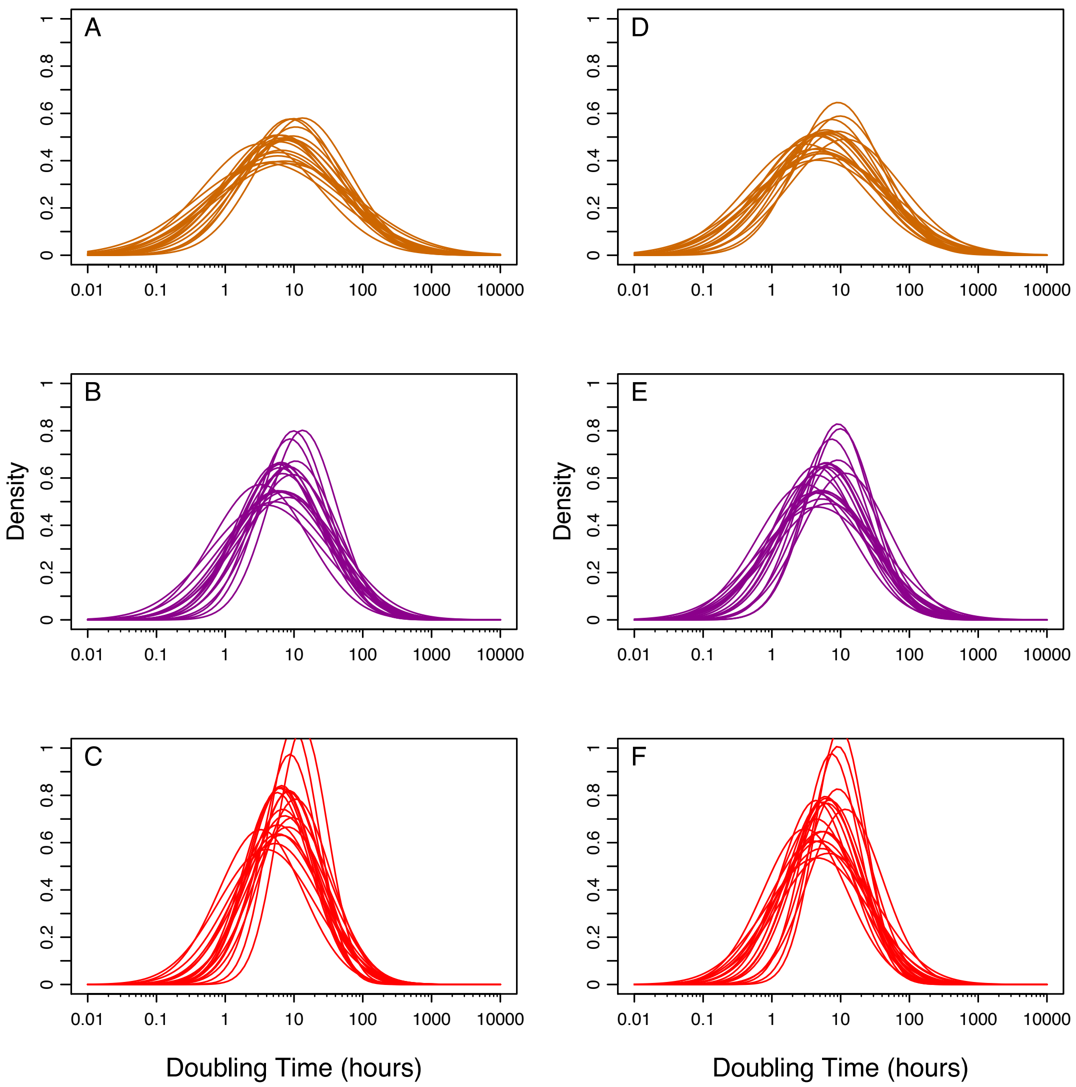


Figure 10. DT distributions inferred by bootstrapping the accumulation and mutation rate data and refitting the log-normal distributions to both datasets. Each plot shows 20 bootstrap DT distributions assuming different levels of correlation between the accumulation and mutation rates - orange r = 0, purple r = 0.5 and red r = 0.75. A, B and C include all mutation rate estimates and D,E, and F show the analysis after removal of the *Mesoplasma florum* mutation rate estimate.

5. DISCUSSION

The rate at which bacteria accumulate mutations over short timeframes of 1 to 1500 years varies by three orders of magnitude. The rate of accumulation must depend on the mutation rate per year and the strength of natural selection, and in turn the mutation rate per year is likely to depend on the mutation rate per generation and the generation time, assuming that at least some mutations are a consequence of replication errors. Potentially, variation in any of these factors - the mutation rate per generation, the generation time and the strength of selection - could be responsible for the variation in the accumulation rate.

Unfortunately, we find no very clear correlate of the accumulation rate; the accumulation rate is significantly correlated to the GC-content and genome size, but neither factor is significant when we control for phylogeny. There is a hint that both lab DT and the effective population size may be important since these emerge as marginally significant in a multiple regression of all factors when we control for phylogeny. The lack of any clear correlate may be a result of the size of our dataset; we have data from just 34 species and many of the accumulation rates are estimated with considerable error. It is likely that the number of data-points will increase considerably over the coming years and a more powerful analysis will be possible.

It has previously been shown that the accumulation rate is correlated to the timeframe over which the accumulation rate is measured (Duchêne *et al*. 2016). This relationship is expected given that deleterious mutations can segregate in a population, but these are ultimately removed from the population. However, in the study of Duchenne *et al*. (2016) the relationship was largely a consequence of two data-points which were sampled over a very long time period, and Duchenne *et al.* excluded datasets in which there was significant increase in the accumulation of mutations with time. This would bias them towards finding a negative correlation between the accumulation rate and sampling time, because bacteria with slow accumulation rates would be excluded if they had been sampled over a short period of time because they wouldn’t show significant evidence of mutation accumulation. We found no evidence of a relationship between the rate of accumulation and sampling time within bacterial species suggesting that sampling time and accumulation rate are not correlated over the time-frames being considered. This is perhaps not surprising because our theoretical analysis suggests that differences in accumulation rate are only likely to be apparent if some bacteria are sampled over very short and very long time frames. The relationship

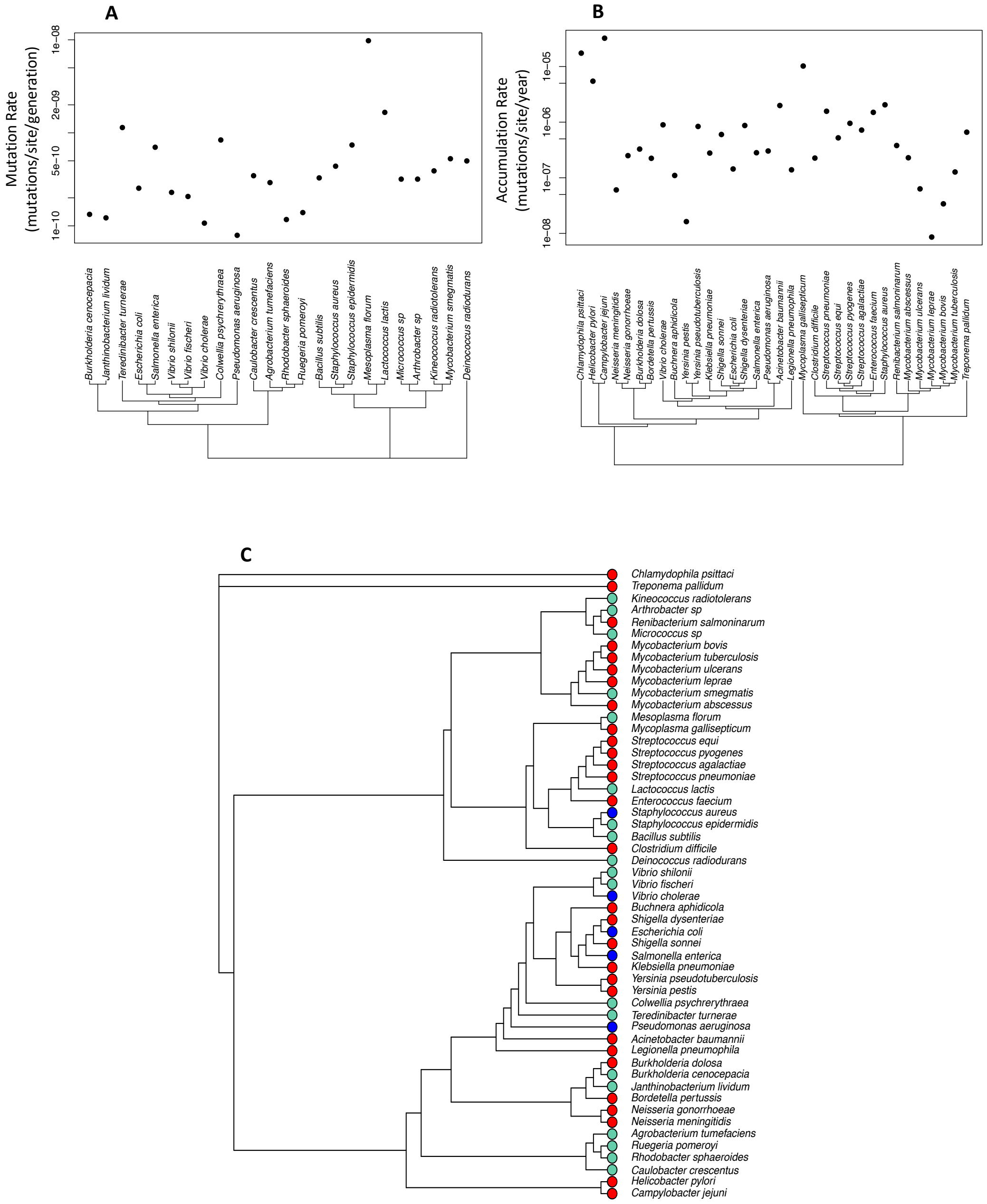


Figure 11. A) 16s rRNA phylogeny and mutation rate estimates for 24 species of bacteria (two species are excluded because of erroneous positioning on the phylogeny - see Figure A4 for details). B) 16s rRNA phylogeny and accumulation rate estimates for 34 species of bacteria. C) 16s rRNA phylogeny combining species for which we have an estimate of the mutation rate and or accumulation rate. Coloured dots indicate which kind of information each species provides - red = accumulation rate, green = mutation rate and blue = both a mutation rate and an accumulation rate.

is very likely to exist but we have been unable to detect it and it is clearly not responsible for most of the variation in the accumulation rate.

We find only very weak evidence that the accumulation rate is correlated to the doubling time, as measured in the lab. However, this is perhaps not surprising. Few bacteria probably double at anything like their lab measured rates in their natural environment. We have recently estimated the DT of 5 bacterial species indirectly. We have used estimates of the rate at which bacteria accumulate mutations in their natural environment and estimates of the rate at which they mutate in the laboratory to estimate the DT for these 5 bacteria and infer the distribution of DTs across bacteria. We estimate that DTs are generally longer in the wild than in the lab, but critically we also infer that DTs vary by several orders of magnitude between bacterial species and that many bacteria have very slow DTs in their natural environment.

The method, by which we have inferred the DT in the wild, makes three important assumptions. We assume that the mutation rate per generation is the same in the lab and in the wild. However, it seems likely that bacteria in the wild will have a higher mutation rate per generation than those in the lab for two reasons. First, bacteria in the wild are likely to be stressed and this can be expected to elevate the mutation rate (Bjedov *et al*. 2003; Galhardo *et al.* 2007; Foster 2007; Maclean *et al.* 2013; Shewaramani *et al.* 2017). Second, if we assume that DTs are longer in the wild than the lab then we expect the mutation rate per generation to be higher in the wild than in the lab because some mutational processes do not depend upon DNA replication. The relative contribution of replication dependent and independent mutational mechanisms to the overall mutation rate is unknown. Rates of substitution are higher in Firmicutes that do not undergo sporulation suggesting that replication is a source of mutations in this group of bacteria (Weller & Wu 2015), but see Maugham (2007). However, rates of mutation accumulation seem to be similar in latent versus active infections of *M. tuberculosis*, suggesting that replication independent mutations might dominate in this bacterium (Ford *et al.* 2011; Lillebaek *et al.* 2016).

The second major assumption is that the rate at which mutations accumulate in the wild is equal to the mutation rate per year; in effect, we are assuming that all mutations are effectively neutral, at least over the timeframe in which they are assayed (or that some are inviable, but the same proportion are inviable in the wild and the lab). In those accumulation rate studies, in which they have been studied separately, non-synonymous mutations accumulate more slowly than synonymous mutations; relative rates vary from 0.13 to 0.8, with a mean of 0.57 (Table A3). There is no correlation between the time-frame over which the estimate was made and the ratio of non-synonymous and synonymous accumulation rates ($r = 0.2$, $p = 0.53$). We did not attempt to control for selection because the relative rates of synonymous and non-synonymous accumulation are only available for a few species, and the relative rates vary between species. However, we can estimate the degree to which more selection against deleterious non-synonymous accumulations in the wild causes the DT to be underestimated as follows. The observed rate at which mutations accumulate in a bacterial lineage is

$$\mu_{obs} = \alpha\, \mu_{true}\, \delta_i + (1-\alpha)(1-\beta)\, \mu_{true}\, \delta_s + (1-\alpha)\, \beta\, \mu_{true}\, \delta_n, \qquad (5)$$

where α is the proportion of the genome that is non-coding and β is the proportion of mutations in protein coding sequence that are non-synonymous. δ_x is the proportion of mutations of class x (i = intergenic, s = synonymous, n = non-synonymous) that are effectively neutral. α and β are approximately 0.15 and 0.7, respectively, in our dataset. Although there is selection on synonymous codon use in many bacteria (Hershberg & Petrov 2008), selection appears to be weak (Sharp *et al.* 2005) we therefore assume that $\delta_s = 1$. This implies, from the rate at which non-synonymous mutations accumulate relative to synonymous mutations, that $\delta_n = 0.6$. A recent analysis of intergenic regions in several species of bacteria has concluded that selection is weaker in intergenic regions than at non-synonymous sites, we therefore assume that $\delta_i = 0.8$ (Thorpe *et al.* 2017). Using these estimates suggests that selection leads us to underestimate the true mutation rate per year in the wild by ~27%; this in turn means we have over-estimated the DT by ~37%, a relatively small effect. To investigate how sensitive this estimate is to the parameters in equation 1, we varied each of them in turn (Table 3). We find that the observed mutation rate is most sensitive to selection on synonymous codon use, because if there is selection on synonymous codon use this also affects our estimates of selection at non-synonymous sites and in intergenic. For example, if selection on synonymous codon use depressed the synonymous accumulation rate by 0.5 this would lead to an underestimate of the mutation rate of 63%, which would in turn have led to a 2.7 fold over-estimate of the DT.

Finally, although each study attempted to remove SNPs that had arisen by recombination, it is possible that some are still present in the data. Recombinant SNPs can have two effects. First, if they have recombined from outside the clade they inflate the accumulation rate estimate and hence lead to an underestimate of the DT. Second, if there is recombination within a clade, they affect the phylogeny and potentially lead to the root of the tree being estimated as younger than it should be. This will lead to an over-estimate of the DT.

It is important to appreciate that our method estimates an average DT within a particular environment that the bacteria were sampled from. The bacterium may go through periods of quiescence interspersed with periods of growth.

Despite the assumptions we have made in our method, our estimate of the DT of *P. aruginosa* of 2.3 hours in a cystic fibrosis patient is very similar to that independently estimated using the ribosomal content of cells of between 1.9 and 2.4 hours (Yang *et al.* 2008). There is also independent evidence that there are some bacteria that divide slowly in their natural environment. The aphid symbiont *Buchnera aphidicola* is estimated to double every 175-292 hours in its host (Ochman *et al.* 1999; Clark *et al.* 1999), and *Mycobacterium leprae* doubles every 300-600 hours on mouse footpads (Shepard 1960; Rees 1964; Levy 1976), not its natural environment, but one that is probably similar to the human skin. Furthermore, in a recent selection experiment, Avrani *et al.* (2017) found that several *E. coli* populations, which were starved of resources, accumulated mutations in the core RNA polymerase gene. These mutations caused these strains to divide more slowly than unmutated strains when resources were plentiful. Interestingly these same mutations are found at high frequency in

δs	**δi**	**δn**	**mu obs/ mu true**	**DT (if DT obs =10hrs)**
1	0.9	0.8	0.87	11.49
""	""	0.6	0.75	13.33
""	""	0.3	0.57	17.54
""	0.8	0.8	0.85	11.76
""	""	0.6	0.73	13.70
""	""	0.3	0.55	18.18
""	0.4	0.8	0.79	12.66
""	""	0.6	0.67	14.93
""	""	0.3	0.49	20.41
0.5	0.45	0.4	0.43	23.26
""	""	0.3	0.37	27.03
""	""	0.15	0.28	35.71
""	0.4	0.4	0.43	23.26
""	""	0.3	0.37	27.03
""	""	0.15	0.28	35.71
""	0.2	0.4	0.4	25.00
""	""	0.3	0.34	29.41
""	""	0.15	0.25	40.00

Table 3. Testing different parameter combinations to investigate how sensitive the doubling time estimate is to the parameters in equation 5. Each parameter is varied in turn. δi and δn are dependent on δs so they are halved when δs is halved.

unculturable bacteria, suggesting that there is a class of slow growing bacteria in the environment that are adapted to starvation.

Korem *et al*. (2015) have recently proposed a general method by which the DT can be potentially estimated. They note that actively replicating bacterial cells have two or more copies of the chromosome near the origin of replication but only one copy near the terminus, if cell division occurs rapidly after the completion of DNA replication. Using next generation sequencing, they show that it is possible to assay this signal and that the ratio of sequencing depth near the origin and terminus is correlated to bacterial growth rates *in vivo*. Brown *et al.* (2016) have extended the method to bacteria without a reference genome and/or those without a known origin and terminus of replication. In principle, these measures of cells performing DNA replication could be used to estimate the DT of bacteria in the wild. However, it's unclear how or whether the methods can be calibrated. Both Korem *et al*. (2015) and Brown *et al.* (2016) find that their replication measures have a median of ~1.3 across bacteria in the human gut. However, a value of 1.3 translates into different relative and absolute values of the DT in the two studies. Brown *et al.* (2016) show that their measure of replication, iRep, is highly correlated to Korem et al.'s (2015) measure, PTR, for data from *Lactobacillus gasseri*; the equation relating the two statistics is iRep = -0.75 + 2 PTR. Hence, when PTR = 1.3, iRep = 1.85 and when iRep = 1.3, PTR = 1.03. The two methods are not consistent. They also yield very different estimates for the absolute DT. Korem *et al.* (2015) show that PTR is highly correlated to the growth rate of *E. coli* grown in a chemostat. If we assume that the relationship between PTR and growth rate is the same across bacteria *in vivo* and *in vitro*, then this implies that the median DT for the human microbiome is ~2.5 hours. In contrast, Brown *et al*. (2016) estimate the growth rate of *Klebsiella oxytoca* to be 19.7 hours in a new-born baby using faecal counts and find that this population has an iRep value of ~1.77. This value is greater than the vast majority of bacteria in the human microbiome and bacteria in the Candidate Phyla Radiation, suggesting that most bacteria in these two communities replicate very slowly. These discrepancies between the two methods suggest that it may not be easy to calibrate the PTR and iRep methods to yield estimates of the DT across bacteria.

How should we interpret our results for the five focal species in the context of what is known of their ecology? *Vibrio cholerae* displays the shortest DT of 1.1 hr. *Vibrio* species are ubiquitous in estuarine and marine environments (Reidl & Klose 2002).

They are known to have very short generation times in culture, the shortest being *V. natriegens* of just 9.8 minutes (G. 1961). In the wild they can exploit a wide range of carbon and energy sources, and as such have been termed “opportunitrophs” (Polz *et al.* 2006). Natural *Vibrio* communities do not grow at an accelerated rate continuously, but can exist for long periods in a semi-dormant state punctuated by rapid pulses of high growth rates (Blokesch & Schoolnik 2008), or blooms (Takemura *et al*. 2014), when conditions are favourable. It has also been argued that the unusual division of *Vibrio* genomes into two chromosomes facilitates more rapid growth (Yamaichi *et al.* 1999). By pointing to a very short DT in *V. cholerae*, our analysis is therefore consistent with what is known of the ecology of this species.

Staphylococcus aureus is predominantly found on animals and humans and inhabits various body parts, including the skin and upper respiratory tract (Schenck *et al.* 2016). It can cause infection of the skin and soft tissue as well bacteraemia (John 2004). *S. aureus* exhibits a range of modes of growth, some of which may to allow it to survive stress and antimicrobials whilst in its host. For instance, small subpopulations can adopt a slow-growing, quasi-dormant lifestyle, either in a multicellular biofilm or as small colony variants (SCVs) or persister cells (Bui *et al.* 2017). Our short DT of 1.8 hours suggests this is not the typical state for *S. aureus* in the wild, which is not surprising considering the incidence of SCVs in clinical samples is fairly low, between 1-30% (Proctor *et al*. 2006).

Pseudomonas aeruginosa can inhabit a wide variety of environments, including soil, water plants and animals. Like our other focal species, it is an opportunistic pathogen and can also infect humans, especially those with compromised immune systems, such as patients with cystic fibrosis (CF). In this context infection is chronic. Parallel evolution, the differential regulation of genes which allow it to evade the host immune system and resist antibiotic treatment during infection (Huse *et al.* 2010), and evidence of positive selection (Smith *et al.* 2006) suggests *P. aeruginosa* can adapt to the lungs of individuals with CF for its long-term survival. It is known to actively grow in sputum (Yang *et al.* 2008), where it utilises the available nutrition which supports its growth to high population densities (Palmer *et al*. 2005). Its ability to adapt and actively grow in the CF sputum is consistent with its relatively short DT of 2.3 hours, especially considering this is the environment in which the accumulation rate was measured and matches that estimated by Yang et al 2008 (Yang *et al*. 2008).

E. coli and *S. enterica* primarily reside in the lower intestine of humans and animals, but can also survive in the environment. Although *E. coli* is commonly recovered from environmental samples, it is not thought able to grow or survive for prolonged periods outside of the guts of warm blooded animals, except in tropical regions where conditions are more favourable (Winfiel & Groisman 2003), although some phylogenetically distinct strains appear to reproduce and survive well in the environment (Oh *et al.* 2012). In contrast, *Salmonella* is also an enteric coloniser of cold-blooded animals, in particular reptiles, is better adapted than *E. coli* at surviving and growing in environmental niches. For example, *Salmonella* can survive and grow for at least a year in soil (Davies & Wray 1996), whereas *E. coli* can survive for only a few days (Bogosian & Sammons 1996). Although these secondary niches may play a greater role in *Salmonella* than in *E. coli*, it remains the case the growth rates in the environment will be much lower than those in a gut. Therefore, the increased tenacity of Salmonella in non-host environments compared to *E. coli* might help to explain the slower DT in this species.

In summary, the availability of accumulation and mutation rate estimates allows us to infer the DT for bacteria in the wild, and the distribution of wild DTs across bacterial species. These DT estimates are likely to be underestimates because the mutation rate per generation is expected to be higher in the wild than in the lab, and some mutations are not generated by DNA replication. Our analysis therefore suggests that DTs in the wild are typically longer than those in the lab, that they vary considerably between bacterial species and that a substantial proportion of species have very long DTs in the wild. This then would explain why accumulation rates vary so widely, there is a very large variance in DTs.

6. CONCLUSION

We wanted to assess the factors that potentially correlate with the accumulation rate in bacteria to investigate whether we could explain the variation in the accumulation rate found across different species. In total we collected accumulation rate estimates for 34 species of bacteria, which were mostly pathogens evolving either within individual hosts or during an outbreak. These estimates varied 3700-fold and the time-frame over which they were measured was between 1-1500 years. There are several factors that could be responsible for this huge variation including the mutation rate, natural selection and the time-frame over which rates are measured. Whilst genome size and GC content, which are proxies for the mutation rate per generation, showed a significant relationship with accumulation rate, after controlling for phylogenetic non-independence this relationship was lost. Similarly, a measurement for the effectiveness of selection, πN/πS, revealed an almost significant correlation to the accumulation rate, which was again lost when we control for phylogeny. No correlation was found between pN/pS for the strains used to estimate the accumulation rate and the accumulation rate.

Surprisingly, we find little evidence that the sampling time correlates with the accumulation rate. We find a significant negative correlation between sampling time and the accumulation rate, however this appears to be mainly driven by two species, *Yersinia pestis* and *Mycobacterium leprae*, which were sampled over relatively long time frames.

One final factor that should influence the accumulation rate is generation time. We find no relationship between lab doubling times and the accumulation rate. However, to further this analysis we developed a method to estimate doubling times in the wild. We estimate this value for five species of bacteria and also the distribution of DTs across all bacteria. Both suggest that DTs for bacteria in the wild are considerably longer than those in the laboratory. Furthermore, they vary by orders of magnitude between different species and it appears that many species double very slowly in the wild. In conclusion, no one factor tested here stands out as a clear candidate for explaining the variation in the accumulation rates of bacteria. We can, however, suggest that due to the large variation seen in bacterial doubling times in the wild this could be the major factor driving the variation in the accumulation rate across species.

7. BIBLIOGRAPHY

Alam, M.T. et al., 2015. Transmission and Microevolution of USA300 MRSA in U.S. Households: Evidence from Whole-Genome Sequencing. *mBio*, 6(2), pp.e00054-15.

Avrani, S. et al., 2017. Rapid Genetic Adaptation during the First Four Months of Survival under Resource Exhaustion. *Molecular Biology and Evolution*, 34(7), pp.1758–1769.

Baines, S.L. et al., 2015. Convergent adaptation in the dominant global hospital clone ST239 of methicillin-resistant Staphylococcus aureus. *mBio*, 6(2), pp.e00080-15.

Balbi, K.J. & Feil, E.J., 2007. The rise and fall of deleterious mutation. *Research in Microbiology*, 158(10), pp.779–786.

Biek, R. et al., 2015. Measurably evolving pathogens in the genomic era. *Trends in Ecology & Evolution*, 30(6), pp.306–313.

Bjedov, I. et al., 2003. Stress-Induced Mutagenesis in Bacteria. *Science*, 300, pp.1404–1409.

Blokesch, M. & Schoolnik, G.K., 2008. The extracellular nuclease Dns and its role in natural transformation of Vibrio cholerae. *Journal of Bacteriology*, 190(21), pp.7232–7240.

Blomberg, S.P., Garland, T. & Ives, A.R., 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution*, 57(4), pp.717–745.

Bogosian, G. & Sammons, L., 1996. Death of the *Escherichia coli* K-12 strain W3110 in soil and water. *Applied and environmental microbiology*, 62(11), pp.4114–4120.

Bromham, L., 2002. Molecular clocks in reptiles: life history influences rate of molecular evolution. *Molecular biology and evolution*, 19(3), pp.302–309.

Brown, C.T. et al., 2016. Measurement of bacterial replication rates in microbial communities. *Nature Biotechnology*, 34(12), pp.1256–1263.

Bui, L.M.G., Conlon, B.P. & Kidd, S.P., 2017. Antibiotic tolerance and the alternative lifestyles of *Staphylococcus aureus*. *Essays In Biochemistry*, 61(1), pp.71–79.

Clark, M. a, Moran, N. a & Baumann, P., 1999. Sequence evolution in bacterial endosymbionts having extreme base compositions. *Molecular biology and evolution*, 16, pp.1586–1598.

Cui, Y. et al., 2013. Historical variations in mutation rate in an epidemic pathogen, Yersinia pestis. *Proceedings of the National Academy of Sciences*, 110(2), pp.577–582.

Davies, R.H. & Wray, C., 1996. Seasonal variations in the isolation of Salmonella typhimurium, Salmonella enteritidis, Bacillus cereus and Clostridium perfringens from environmental samples. *Zentralblatt fur Veterinarmedizin. Reihe B. Journal of veterinary medicine. Series B*, 43(2), p.119—127.

Dettman, J.R., Sztepanacz, J.L. & Kassen, R., 2016. The properties of spontaneous mutations in the opportunistic pathogen Pseudomonas aeruginosa. *BMC Genomics*, pp.1–14.

Drake, J.W., 1991. A constant rate of spontaneous mutation in DNA-based microbes. *Proceedings of the National Academy of Sciences*, 88(August), pp.7160–7164.

Drummond, A.J. et al., 2003. Measurably evolving populations. *Trends in Ecology & Evolution*, 18(9), pp.481–488.

Duchêne, S. et al., 2016. Genome-scale rates of evolutionary change in bacteria. *Microbial Genomics*, 2(October 2016).

Duchene, 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.

Edgar, R.C., 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5), pp.1792–1797.

Felsenstein, J., 1985. Phylogenies and the Comparative Method. *The American Naturalist*, 125(1), pp.1–15.

Ford, C.B. et al., 2011. Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection. *Nature Genetics*, 43(5), pp.482–486.

Foster, P.L., 2007. Stress-induced mutagenesis in bacteria. *Critical reviews in biochemistry and molecular biology*, 42, pp.373–97. Available at:

G., E.R., 1961. Pseudomonas natriegens, a marine bacterium with a generation time of less than 10 minutes. *Journal of bacteriology*, 83, pp.736–737.

Galhardo, R.S., Hastings, P.J. & Rosenberg, S.M., 2007. Mutation as a stress response and the regulation of evolvability, *Critical Reviews in Biochemistry and Molecular Biology*, 42, pp.399–435.

Gibbons, R.J. & Kapsimalis, B., 1967. Estimates of the overall rate of growth of the intestinal microflora of hamsters, guinea pigs, and mice. *Journal of Bacteriology*, 93(1), pp.510–512.

Gray, R.R. et al., 2011. Testing spatiotemporal hypothesis of bacterial evolution using methicillin-resistant staphylococcus aureus ST239 genome-wide data within a bayesian framework. *Molecular Biology and Evolution*, 28(5), pp.1593–1603.

Harmsen, H.J.M. et al., 1998. Syntrophobacter furnaroxidans sp. nov., a syntrophic propionate-degrading sulfate- reducing bacterium. *International Journal of Systematic Bacteriology*, 48, pp.1383–1388.

Harris, S.R. et al., 2010. Evolution of MRSA During Hospital Transmission and Intercontinental Spread. *Science*, 327, pp.469–474.

Hawkey, J. et al., 2013. Evidence of microevolution of Salmonella Typhimurium during a series of egg-associated outbreaks linked to a single chicken farm. *BMC genomics*, 14(1), pp.800.

Hershberg, R. & Petrov, D.A., 2010. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genetics*, 6(9), pp.e1001115.

Hershberg, R. & Petrov, D.A., 2008. Selection on Codon Bias. *Annual review of Genetics*, 42, pp.287–299.

Hildebrand, F., Meyer, A. & Eyre-Walker, A., 2010. Evidence of selection upon genomic GC-content in bacteria. *PLoS Genetics*, 6(9), pp.e1001107.

Ho, S.Y.W. et al., 2011. Time-dependent rates of molecular evolution. *Molecular Ecology*, 20(15), pp.3087–3101.

Ho, S.Y.W. & Larson, G., 2006. Molecular clocks: When timesare a-changin'. *Trends in Genetics*, 22(2), pp.79–83.

Holden, M.T.G. et al., 2013. A genomic portrait of the emerfences, evolution, and global spread of methicillin-resistant Staphylococcus aureus. *Genome research*, 23, pp.653–664.

Hughes, A.L., 2005. Evidence for Abundant Slightly Deleterious Polymorphisms in Bacterial Populations. , 538(February), pp.533–538.

Huse, H.K. et al., 2010. Parallel evolution in *Pseudomonas aeruginosa* over 39,000 generations *in vivo*. *mBio*, 1(4), pp.e00199-10.

John, J.F., 2004. Staphylococcal Infection: Emerging Clinical Syndromes, *Horwood*

Publishing Limited, pp 1-30*.*

Kearse, M. et al., 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12), pp.1647–1649.

Kennemann, L. et al., 2011. Helicobacter pylori genome evolution during human infection. *Proceedings of the National Academy of Sciences of the United States of America*, 108(12), pp.5033–5038.

Korem, T. et al., 2015. Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science*, 349(6252), pp.1101–1106.

Krasovec, M. et al., 2017. Spontaneous mutation rate in the smallest photosynthetic eukaryotes. *Molecular biology and evolution*, 34(7), pp.1–26.

Lanfear, R., Welch, J.J. & Bromham, L., 2010. Watching the clock: Studying variation in rates of molecular evolution between species. *Trends in Ecology and Evolution*, 25(9), pp.495–503.

Lassalle, F. et al., 2015. GC-Content evolution in bacterial genomes: the biased gene conversion hypothesis expands. *PLoS genetics*, 11(2), pp.e1004941.

Lee, H. et al., 2012. PNAS Plus: Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *Proceedings of the National Academy of Sciences*, 109(41), pp.E2774–E2783.

Levy, L., 1976. Bactericidal action of dapsone against Mycobacterium leprae in mice. *Antimicrobial Agents and Chemotherapy*, 9(4), pp.614–617.

Lillebaek, T. et al., 2016. Substantial molecular evolution and mutation rates in prolonged latent Mycobacterium tuberculosis infection in humans. *International Journal of Medical Microbiology*, 306(7), pp.580–585.

Lind, P. a & Andersson, D.I., 2008. Whole-genome mutational biases in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 105(46), pp.17878–17883.

Long, H. et al., 2015. Background mutational features of the radiation-resistant bacterium Deinococcus radiodurans. *Molecular biology and evolution*, 32(9), pp.2383–2392.

Long, H. et al., 2018. Evolutionary determinants of genome-wide nucleotide composition. *Nature Ecology and Evolution*, 2(February), pp.1–4.

Lynch, M., 2010. Evolution of the mutation rate. *Trends in Genetics*, 26(8), pp.345–352.

Lynch, M., Ackerman, M.S., et al., 2016. Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics*, 17(11), pp.704–714.

Lynch, M., Ackerman, M.S., et al., 2016. Genetic drift , selection and the evolution of the mutation rate. *Nature Publishing Group*, 17(11), pp.704–714.

Maclean, R.C., Torres-barceló, C. & Moxon, R., 2013. Evaluating evolutionary models of stress-induced mutagenesis in bacteria. *Nature Reviews Genetics*, 14(3), pp.221–227.

Markussen, T. & Marvig, R.L., 2014. Environmental Heterogeneity Drives Within-Host Diversification and Evolution of Pseudomonas aeruginosa. *mBio*, 5(5), pp.e01592-14.

Marvig, R.L. et al., 2013. Genome analysis of a transmissible lineage of pseudomonas aeruginosa reveals pathoadaptive mutations and distinct evolutionary paths of hypermutators. *PLoS genetics*, 9(9), pp.e1003741.

Mather, A.E., 2013. Distinguishable Epidemics of Multidrug-Resistant Salmonella Typhimurium DT104 in Different Hosts. *Science*, 341, pp.1514–1518.

Maughan, H., 2007. Rates of Molecular Evolution in Bacteria are Relatively Constant Despite Spore Dormancy. *Evolution, February,* pp.280-288

Mutreja, A. et al., 2011. Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature*, 477, pp.462–465.

Nübel, U. et al., 2010. A Timescale for Evolution, Population Expansion, and Spatial Spread of an Emerging Clone of Methicillin-Resistant Staphylococcus aureus. *PLoS Pathogens*, 6(4), p.e1000855.

Ochman, H., Elwyn, S. & Moran, N.A., 1999. Calibrating bacterial evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 96(22), pp.12638–43.

Oh, S. et al., 2012. Genomic Diversity of Escherichia Isolates from Diverse Habitats. *PLoS ONE*, 7(10), pp.e47005.

Okoro, C.K. et al., 2012. Intracontinental spread of human invasive Salmonella Typhimurium pathovariants in sub-Saharan Africa. *Nature Genetics*, 44(11), pp.1215–1221.

Pagel, M., 1999. Inferring the historical patterns of biological evolution. *Nature*, 401(6756), pp.877–884.

Palmer, K.L. et al., 2005. Cystic Fibrosis Sputum Supports Growth and Cues Key Aspects of Pseudomonas aeruginosa Physiology Cystic Fibrosis Sputum Supports Growth and Cues Key Aspects of Pseudomonas aeruginosa Physiology. *Journal of Bacteriology*, 187(15), pp.5267–5277.

Paradis, E., Claude, J. & Strimmer, K., 2004. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2), pp.289–290.

Polz, M.F. et al., 2006. Patterns and mechanisms of genetic and phenotypic differentiation in marine microbes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1475), pp.2009–2021.

Proctor, R.A. et al., 2006. Small colony variants: A pathogenic form of bacteria that facilitates persistent and recurrent infections. *Nature Reviews Microbiology*, 4(4), pp.295–305.

Rees, R.J., 1964. Limited Multiplication of Acid-Fast Bacilli in the Foot-Pads of Mice Inoculated With Mycobacterium Leprae. *British journal of experimental pathology*, 45(2), pp.207–18.

Reeves, P.R. et al., 2011. Rates of Mutation and Host Transmission for an Escherichia coli Clone over 3 Years. *PloS One*, 6(10), p.e26907.

Reidl, J. & Klose, K.E., 2002. Vibrio cholerae and cholera:out of the water and into the host. *FEMS Microbiology Reviews*, 26(October), pp.125–139.

Revell, L.J., 2012. phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3(2), pp.217–223.

Rocha, E.P. et al., 2006. Comparisons of dN/dS are time dependent for closely related bacterial genomes. *Journal of theoretical biology*, 239.

Schenck, L.P., Surette, M.G. & Bowdish, D.M.E., 2016. Composition and immunological significance of the upper respiratory tract microbiota. *FEBS Letters*, 590(21), pp.3705–3720.

Sharp, P.M. et al., 2005. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Research*, 33(4), pp.1141–1153.

Shepard, C.C., 1960. The Experimental Disease That Follows the Injection of Human Leprosy Bacilli Into Foot-Pads of Mice. *The Journal of experimental medicine*, 112(3), pp.445–54.

Shewaramani, S. et al., 2017. Anaerobically Grown Escherichia coli Has an Enhanced Mutation Rate and Distinct Mutational Spectra. , 13, pp.1–22.

Smith, E.E. et al., 2006. Genetic adaptation by Pseudomonas aeruginosa to the airways of cystic fibrosis patients. *Proceedings of the National Academy of Sciences of the United States of America*, 103(22), pp.8487–92.

Smith, S.A. & Donoghue, M.J., 2008. Rates of molecular evolution are linked to life history in flowering plants. *Science*, 322(5898), pp.86–89.

Smyth, D.S. et al., 2010. Population structure of a hybrid clonal group of methicillin-resistant Staphylococcus aureus, ST239-MRSA-III. *PLoS ONE*, 5(1), pp.e8582.

Stamatakis, A., 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9), pp.1312–1313.

Stinear, T.P. et al., 2014. Adaptive change inferred from genomic population analysis of the ST93 epidemic clone of community-associated methicillin-resistant Staphylococcus aureus. *Genome Biology and Evolution*, 6(2), pp.366–378.

Sun, Y. et al., 2017. Spontaneous mutations of a model heterotrophic marine bacterium. *ISME Journal*, 11(7), pp.1713–1718.

Sung, W. et al., 2012. Drift-barrier hypothesis and mutation-rate evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 109, pp.18488–92.

Takemura, A.F., Chien, D.M. & Polz, M.F., 2014. Associations and dynamics of vibrionaceae in the environment, from the genus to the population level. *Frontiers in Microbiology*, 5(FEB), pp.1–26.

Thorpe, H.A. et al., 2017. Comparative Analyses of Selection Operating on Non-translated Intergenic Regions of Diverse Bacterial Species. *Genetics*, 206, pp.363–376.

Uhlemann, A.-C. et al., 2014. Molecular tracing of the emergence, diversification, and transmission of S. aureus sequence type 8 in a New York community. *Proceedings of the National Academy of Sciences*, 111(18), pp.6738–6743.

Vieira-Silva, S. & Rocha, E.P.C., 2010. The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genetics*, 6(1), p.e1000808.

Vos, M. et al., 2013. ODoSE : A Webserver for Genome-Wide Calculation of Adaptive

Divergence in Prokaryotes. , 8(5), pp.8–11.

Ward, M.J. et al., 2014. Time-scaled evolutionary analysis of the transmission and antibiotic resistance dynamics of Staphylococcus aureus clonal complex 398. *Applied and Environmental Microbiology*, 80(23), pp.7275–7282.

Welch, J.J., Bininda-Emonds, O.R.P. & Bromham, L., 2008. Correlates of substitution rate variation in mammalian protein-coding sequences. *BMC evolutionary biology*, 8(1), p.53.

Weller, C. & Wu, M., 2015. A generation-time effect on the rate of molecular evolution in bacteria. *Evolution*, 69(3), pp.643–652.

Winfiel, M.D. & Groisman, E.A., 2003. Role of Nonhost Enviroments in the Lifestyles of Salmonella and E. coli. *Applied and Environmental Mcrobiology*, 69(7), pp.3687–3694.

Yamaichi, Y., Iida, T. & Park, K., 1999. Physical and genetic map of the genome of Vibrio. *Molecular Microbiology* , 31(5), pp.1513–1521.

Yang, L. et al., 2008. In situ growth rates and biofilm development of Pseudomonas aeruginosa populations in chronic lung infections. *Journal of bacteriology*, 190(8), pp.2767–76.

Young, B.C. et al., 2012. Evolutionary dynamics of Staphylococcus aureus during progression from carriage to disease. *Proceedings of the National Academy of Sciences of the United States of America*, 109(12), pp.4550–4555.

Zhou, Z. et al., 2014. Transient Darwinian selection in Salmonella enterica serovar Paratyphi A during 450 years of global spread of enteric fever. *Proceedings of the National Academy of Sciences*, 111(33), pp.12199–12204.

8. APPENDIX

8.1 Appendix 1. 81 estimates of the rate at which bacteria accumulate mutations per site per year (the accumulation rate) for 34 species of bacteria.

Species	subgroup	Accumulation Rate (x10^{-7})	Included/ excluded/ recalculated	Reason for exclusion/recalculation	Reference
Acinetobacter baumannii	GC1	15.00	Included		(Holt et al. 2016)
Acinetobacter baumannii	GC2	24.70	Included		(Schultz et al. 2016)
Bordetella pertussis		2.24	Included		(Bart et al. 2014)
Buchnera aphidicola		1.10	Included		(Moran et al. 2009)
Buchnera aphidicola		0.09	Excluded	Very long divergence (50 million years)	(Tamas et al. 2002)
Burkholderia dolosa		3.28	Included		(Lieberman et al. 2011)
Campylobacter jejuni		323.00	Included		(Wilson et al. 2009)
Chlamydia psittaci		174.00	Included		(Read et al. 2013)
Clostridium difficile		3.20	Included		(Didelot et al. 2012)
Clostridium difficile	27	1.70	Included		(Steglich et al. 2015)
Clostridium difficile	027/BI/NAP1	1.88	Included		(He et al. 2013)
Enterococcus faecium	ST17/ST252	15.00	Included		(Howden et al. 2013)
Escherichia coli		1.44	Recalculated	Unsure about the rationale related to timepoints used in the calculation. We recalculated by running SNP alignment through BEAST	(Reeves et al. 2011)
Helicobacter pylori		410.00	Excluded	SNPs might be recombinant	(Falush et al. 2001)
Helicobacter pylori		29.35	Excluded	Upper limit on estimate of the divergence time is arbitrary	(Morelli, Didelot, et al. 2010)
Helicobacter pylori		138.00	Excluded	Synonymous rate	(Didelot et al. 2013)
Helicobacter pylori		54.5	Recalculated	Cannot be sure that 3yr isolates are a direct descendant of 0yr isolates	(Kennemann et al. 2011)
Klebsiella pneumoniae	CC258 Clade1	2.56	Included		(Duchêne et al. 2016)
Klebsiella pneumoniae	CC258 Clade2	2.99	Included		(Duchêne et al. 2016)
Legionella pneumophilia		1.39	Included		(Sánchez-Busó et al. 2014)
Mycobacterium abscessus	subsp abscessus	3.63	Included		(Bryant et al. 2013)
Mycobacterium abscessus	subsp massiliense	0.95	Included		(Bryant et al. 2013)
Mycobacterium bovis		0.34	Included		(Biek et al. 2012)
Mycobacterium leprae		0.09	Included		(Schuenemann et al. 2013)
Mycobacterium tuberculosis		0.49	Included		(Bos et al. 2014)
Mycobacterium tuberculosis		1.80	Included		(Ford et al. 2011)
Mycobacterium tuberculosis		1.14	Included		(Walker et al. 2013)

Mycobacterium tuberculosis		1.93	Included		(Duchêne et al. 2016)
Mycobacterium tuberculosis		1.00	Included		(Roetzer et al. 2013)
Mycobacterium ulcerans		0.63	Included		(Vandelannoote et al. 2017)
Mycoplasma gallisepticum		102.00	Included		(Delaney et al. 2012)
Neisseria gonorrhoeae		2.50	Included		(Grad et al. 2014)
Neisseria meningitidis		0.61	Included		(Duchêne et al. 2016)
Pseudomonas aeruginosa	DK2	3.95	Included		(Marvig et al. 2013)
Pseudomonas aeruginosa	DK1	2.11	Included		(Markussen & Marvig 2014)
Pseudomonas aeruginosa	DK2	4.30	Excluded	Synonymous rate	(Yang et al. 2011)
Pseudomonas aeruginosa		154.50	Excluded	Hypermutator strains	(Feliziani et al. 2014)
Renibacterium salmoninarum		3.80	Included		(Brynildsrud et al. 2014)
Salmonella enterica	Kentucky	5.35	Included		(Duchêne et al. 2016)
Salmonella enterica	Typhi H58	1.78	Included		(Duchêne et al. 2016)
Salmonella enterica	paratyphi A	1.94	Included		(Zhou et al. 2014)
Salmonella enterica	Agona	0.93	Included		(Zhou et al. 2013)
Salmonella enterica	Typhimurium Lineage II	1.90	Included		(Okoro et al. 2012)
Salmonella enterica	Typhimurium Lineage I	3.90	Included		(Okoro et al. 2012)
Salmonella enterica	Typhimurium	3.35	Included		(Hawkey et al. 2013)
Salmonella enterica	Typhimurium	3.40	Included		(Mather 2013)
Salmonella enterica	Enteritidis	100.00	Excluded	Hypermutator strain	(Klemm et al. 2016)
Shigella dysenteriae	Sd1	8.70	Included		(Njamkepo et al. 2016)
Shigella sonnei		6.00	Included		(Holt et al. 2012)
Staphylococcus aureus	ST93	4.50	Included		(Stinear et al. 2014)
Staphylococcus aureus	ST239	16.00	Included		(Baines et al. 2015)
Staphylococcus aureus	CC398	16.80	Included		(Ward et al. 2014)
Staphylococcus aureus	ST22	13.00	Included		(Holden et al. 2013)
Staphylococcus aureus	USA300	12.20	Included		(Uhlemann et al. 2014)
Staphylococcus aureus	USA300	12.50	Included		(Alam et al. 2015)
Staphylococcus aureus	ST239	32.50	Included		(Smyth et al. 2010)
Staphylococcus aureus	ST239	33.00	Included		(Harris et al. 2010)
Staphylococcus aureus	ST239	37.90	Included		(Gray et al. 2011)
Staphylococcus aureus	ST225	20.00	Included		(Nübel et al. 2010)
Staphylococcus aureus	MSSA	27.20	Included		(Young et al. 2012)
Streptococcus agalactiae	CC1	6.40	Included		(Da Cunha et al. 2014)
Streptococcus agalactiae	CC17	5.60	Included		(Da Cunha et al. 2014)

Streptococcus agalactiae	CC19	9.30	Included		(Da Cunha et al. 2014)
Streptococcus agalactiae	CC23	7.50	Included		(Da Cunha et al. 2014)
Streptococcus equi		5.22	Included		(Harris et al. 2015)
Streptococcus pneumoniae	PMEN1	15.70	Included		(Croucher et al. 2011)
Streptococcus pyogenes	Emm M1	8.06	Included		(Nasser et al. 2014)
Streptococcus pyogenes	emm12	11.00	Included		(Davies et al. 2015)
Treponema pallidum		6.60	Included		(Arora et al. 2016)
Vibrio cholerae		9.60	Included		(Duchêne et al. 2016)
Vibrio cholerae		8.30	Included		(Mutreja et al. 2011)
Vibrio cholerae		2.35	Excluded	Synonymous rate	(Feng et al. 2008)
Yersinia pestis		0.07	Included		(Morelli, Song, et al. 2010)
Yersinia pestis		0.20	Included		(Bos et al. 2011)
Yersinia pestis		0.16	Included		(Duchêne et al. 2016)
Yersinia pestis		0.23	Included		(Duchêne et al. 2016)
Yersinia pseudotuberculosis	ST19	3.87	Included		(Williamson et al. 2017)
Yersinia pseudotuberculosis	ST43	5.63	Included		(Williamson et al. 2017)
Yersinia pseudotuberculosis	ST9	20.10	Included		(Williamson et al. 2017)
Yersinia pseudotuberculosis	ST42	3.57	Included		(Williamson et al. 2017)
Yersinia pseudotuberculosis	ST14	8.67	Included		(Williamson et al. 2017)

8.2 Appendix 2. Species trait data, including Accumulation Rate, Genome size, GC content, Lab Doubling Time and πN/πS.

Species	**Accumulation Rate (x10^{-7})**	**Genome Size (Mb)**	**GC Content (%)**	**Lab Doubling Time (Hours)**	**πN/πS**
Acinetobacter baumannii	19.90	4.036992	39	NA	0.0485
Bordetella pertussis	2.24	4.115152	68	3.8	0.4604
Buchnera aphidicola	1.10	0.591579	25	NA	0.0539
Burkholderia dolosa	3.28	6.409090	67	1.7	NA
Campylobacter jejuni	323.00	1.676753	30	1.5	0.0855
Chlamydophila psittaci	174.00	1.169811	39	2	0.1631
Clostridium difficile	2.26	4.218256	29	1.1	0.1002
Enterococcus faecium	15.00	3.014847	38	NA	0.0772
Escherichia coli	1.44	5.094524	51	0.5	0.0399
Helicobacter pylori	54.50	1.625146	39	2.4	0.0413
Klebsiella pneumoniae	2.78	5.634122	57	NA	0.0688
Legionella pneumophila	1.39	3.430028	38	3.3	0.0959
Mycobacterium abscessus	2.29	5.029509	64	4.5	0.0864
Mycobacterium bovis	0.34	4.360061	66	NA	0.5790
Mycobacterium leprae	0.09	3.268135	58	NA	NA
Mycobacterium tuberculosis	1.27	4.404328	66	19	0.6491
Mycobacterium ulcerans	0.63	5.805760	66		NA
Mycoplasma gallisepticum	102.00	0.969961	32	1	0.1221
Neisseria gonorrhoeae	2.50	2.210647	52	0.58	0.2409
Neisseria meningitidis	0.61	2.189071	52	0.72	0.1070
Pseudomonas aeruginosa	3.03	6.619300	66	0.5	0.1052
Renibacterium salmoninarum	3.80	3.155250	56	24	NA
Salmonella enterica	2.82	4.818012	52	0.4	0.0585
Shigella dysenteriae	8.70	4.520555	51	NA	0.4294
Shigella sonnei	6.00	5.099185	51	0.53	0.4655
Staphylococcus aureus	20.50	2.853610	33	0.4	0.0834
Streptococcus agalactiae	7.20	2.067505	36	1.8	0.1185
Streptococcus equi	5.22	2.140494	42	2.1	0.1043
Streptococcus pneumoniae	15.70	2.115491	40	0.5	0.1117
Streptococcus pyogenes	9.53	1.836517	39	0.4	0.1195
Treponema pallidum	6.60	1.138605	53		NA
Vibrio cholerae	8.95	4.104331	47	0.2	0.0687
Yersinia pestis	0.16	4.749424	48	1.25	0.6856
Yersinia pseudotuberculosis	8.37	4.783753	47	0.5	0.1216

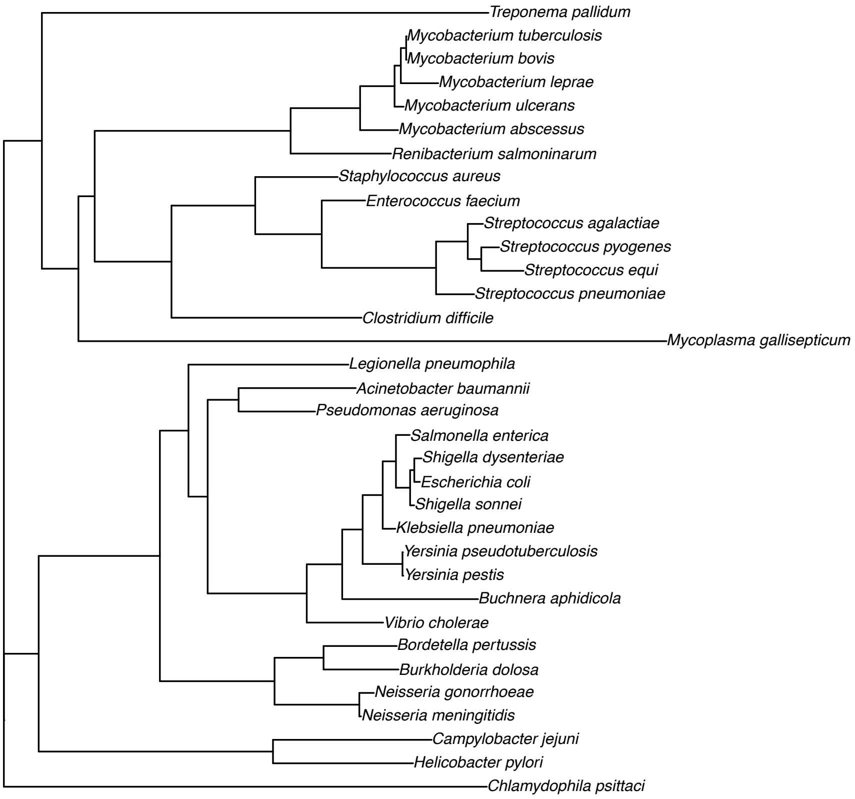
8.3 Appendix 3. Mutation rate per site per generation estimates from Mutation accumulation with whole genome sequencing experiments for 26 species of bacteria.

Species	**Mutation rate/site/generation ($x10^{-10}$)**	**Reference**
Agrobacterium tumefaciens	2.92	(Sung et al. 2016)
Arthrobacter sp	3.18	(Long et al. 2018)
Bacillus subtilis	3.28	(Sung et al. 2015)
Burkholderia cenocepacia	1.33	(Dillon et al. 2015)
Caulobacter crescentus	3.46	(Long et al. 2018)
Colwellia psychrerythraea	8.38	(Long et al. 2018
Deinococcus radiodurans	4.99	(Long et al. 2015)
Escherichia coli	2.54	(Long et al. 2016)
Flavobacterium sp	3.91	(Long et al. 2018
Gemmata obscuriglobus	2.38	(Long et al. 2018)
Janthinobacterium lividum	1.22	(Long et al. 2018)
Kineococcus radiotolerans	3.9	(Long et al. 2018)
Lactococcus lactis	16.6	(Long et al. 2018)
Mesoplasma florum	97.8	(Sung et al. 2012)
Micrococcus sp	3.18	(Long et al. 2018)
Mycobacterium smegmatis	5.27	(Kucukyildirim et al. 2016)
Pseudomonas aeruginosa	0.792	(Dettman et al. 2016)
Rhodobacter sphaeroides	1.17	(Long et al. 2018)
Ruegeria pomeroyi	1.39	(Sun et al. 2017)
Salmonella enterica	7	(Lind & Andersson 2008)
Staphylococcus aureus	4.38	(Long et al. 2018)
Staphylococcus epidermidis	7.4	(Sung et al. 2016)
Teredinibacter turnerae	11.4	(Senra et al. 2018)
Vibrio cholerae	1.07	(Dillon et al. 2016)
Vibrio fischeri	2.07	(Dillon et al. 2016)
Vibrio shilonii	2.29	(Strauss et al. 2017)

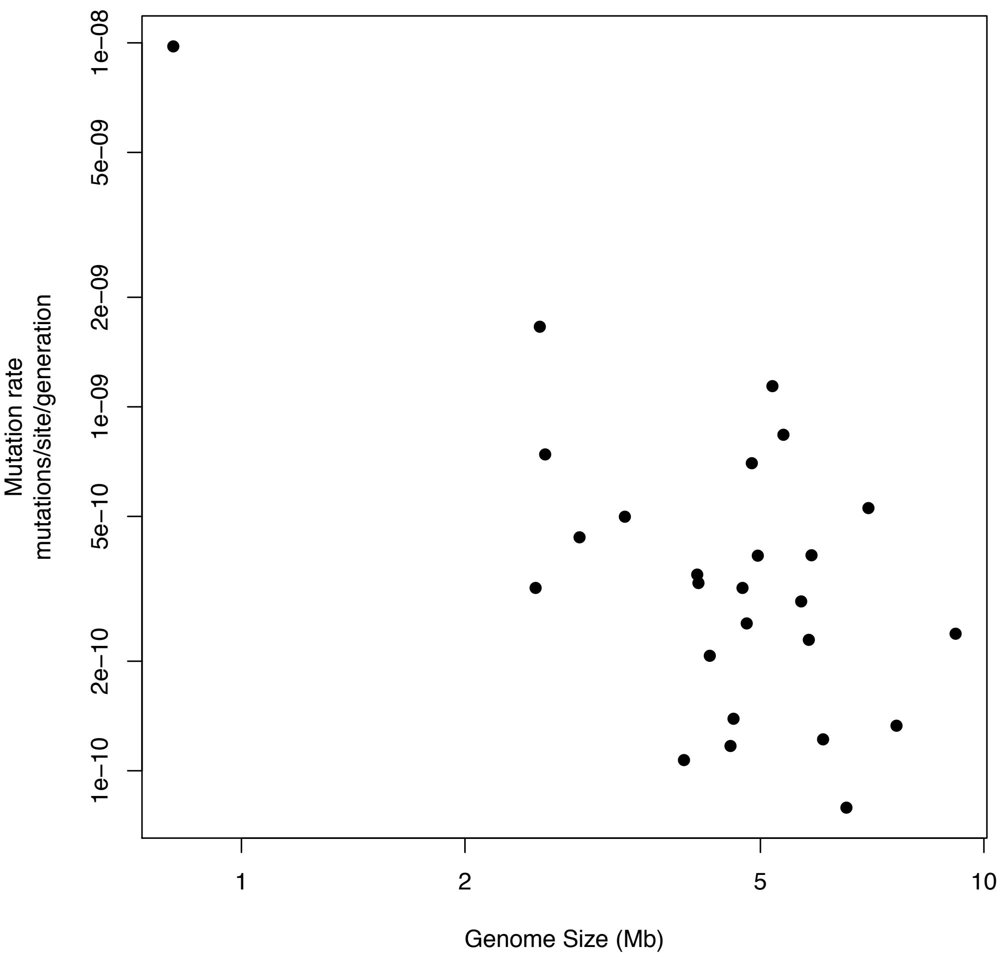
8.4 Appendix 4. dN/dS values for 8 species of bacteria

Species	**dN/dS**	**Reference**
Buchnera aphidicola	0.125	(Moran et al. 2009)
Burkholderia dolosa	1	(Lieberman et al. 2011)
Helicobacter pylori	0.14	(Didelot et al. 2013)
Mycoplasma gallisepticum	0.2	(Delaney et al. 2012)
Pseudomonas aeruginosa DK1	0.56	(Markussen & Marvig 2014)
Pseudomonas aeruginosa DK2	0.66	(Marvig et al. 2013)
Pseudomonas aeruginosa DK2	0.79	(Yang et al. 2011)
Salmonella enterica Agona	0.67	(Zhou et al. 2013)
Salmonella enterica Paratyphi A	0.8	(Zhou et al. 2014)
Salmonella enterica Typhimurium	0.52	(Hawkey et al. 2013)
Staphylococcus aureus (ST225)	0.77	(Nübel et al. 2010)
Streptococcus equi	0.6	(Harris et al. 2015)

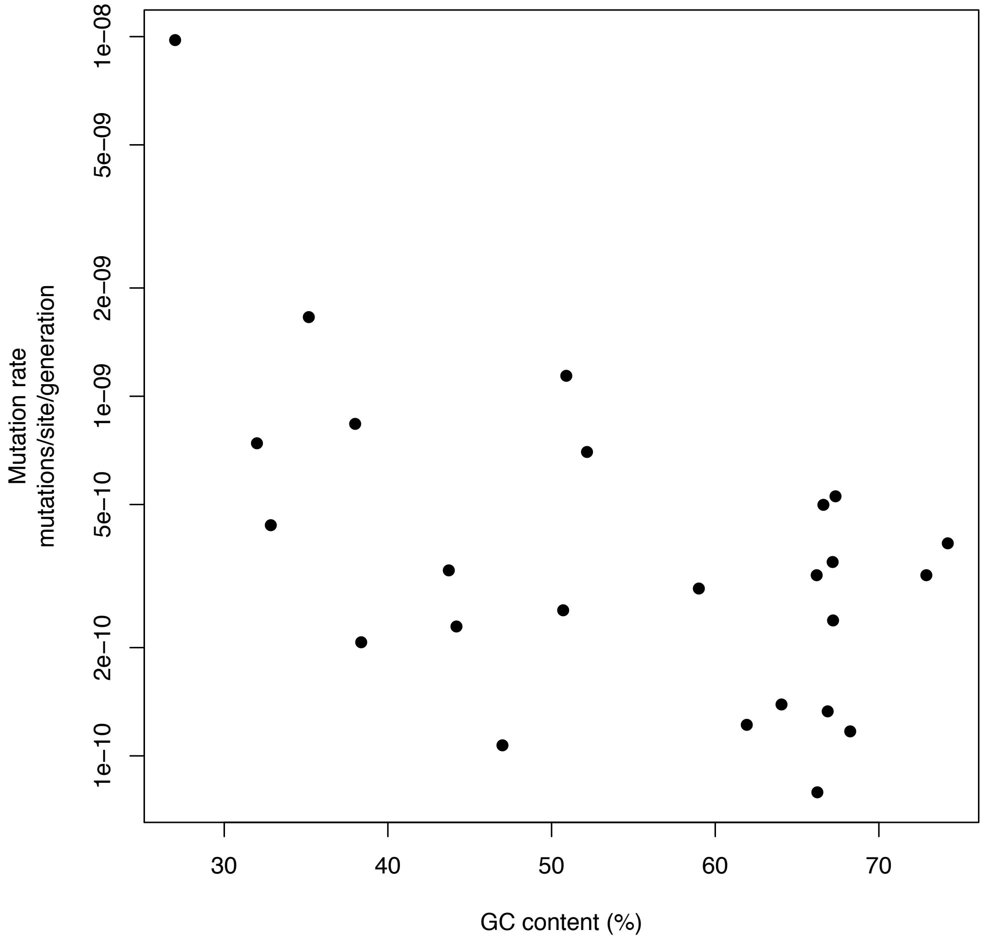
8.5 Appendix 5. 16s rRNA tree for the 34 species of bacteria for which we have an accumulation rate. The tree was used in phylogenetic analyses.



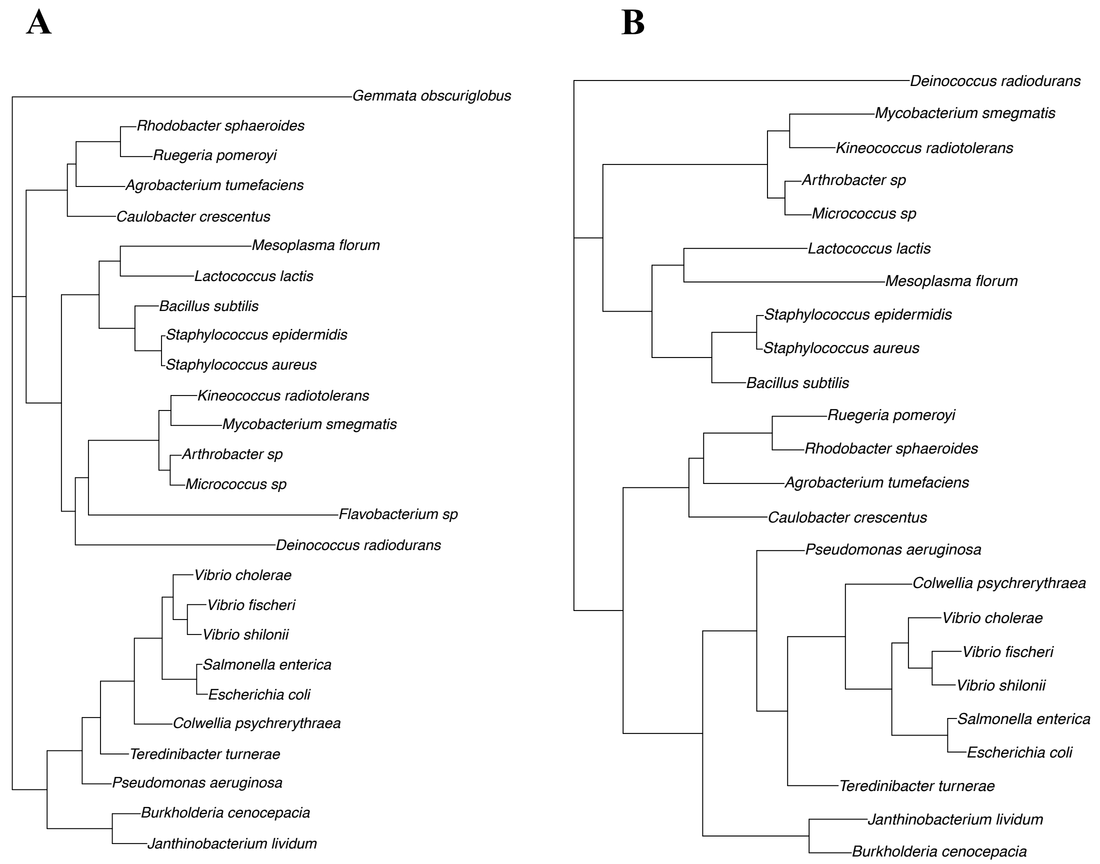
8.6 Appendix 6. The mutation rate/site/generation vs genome size for 26 species of bacteria.



8.7 Appendix 7. The mutation rate/site/generation vs GC content for 26 species of bacteria.



8.8 Appendix 8. 16s rRNA phylogenies for species for which we have a mutation rate estimate. When all 26 species are included for the mutation rate data (A) Flavobacterium sp and the Alphaproteobacteria are erroneously positioned with the gram positive bacteria. This is resolved after exclusion of Flavobacterium sp and Gemmata obscuriglobus. (B).



8.9 Appendix Bibliography

Alam, M.T. et al., 2015. Transmission and Microevolution of USA300 MRSA in U.S. Households: Evidence from Whole-Genome Sequencing. *mBio*, 6(2), pp.e00054-15.

Arora, N. et al., 2016. Origin of modern syphilis and emergence of a contemporary pandemic cluster. *Nature Micro*, 2(December), p.e051037.

Baines, S.L. et al., 2015. Convergent adaptation in the dominant global hospital clone ST239 of methicillin-resistant Staphylococcus aureus. *mBio*, 6(2), pp.e00080-15.

Bart, M.J. et al., 2014. Global Population Structure and Evolution of Bordetella pertussis and Their Relationship with Vaccination. *mBio*, 5(2), pp.e01074-14.

Biek, R. et al., 2012. Whole Genome Sequencing Reveals Local Transmission Patterns of Mycobacterium bovis in Sympatric Cattle and Badger Populations. *PLoS Pathogens*, 8(11), p.e1003008.

Bos, K.I. et al., 2011. A draft genome of Yersinia pestis from victims of the Black Death. *Nature*, 478, pp.506–510.

Bos, K.I. et al., 2014. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature*, 514(7523), pp.494–497.

Bryant, J.M. et al., 2013. Whole-genome sequencing to identify transmission of Mycobacterium abscessus between patients with cystic fibrosis: a retrospective cohort study. *The Lancet*, 381(5), pp.1551–1560.

Brynildsrud, O. et al., 2014. Microevolution of Renibacterium salmoninarum: evidence for intercontinental dissemination associated with fish movements. *The ISME Journal*, 8, pp.746–56.

Croucher, N.J. et al., 2011. Rapid pneumococcal evolution in response to clinical interventions. *Science*, 331, pp.430–4.

Da Cunha, V. et al., 2014. Streptococcus agalactiae clones infecting humans were selected and fixed through the extensive use of tetracycline. *Nature communications*, 5, pp.4544.

Davies, M.R. et al., 2015. Emergence of scarlet fever Streptococcus pyogenes emm12 clones in Hong Kong is associated with toxin acquisition and multidrug resistance. *Nature genetics*, 47(1), pp.84–7.

Delaney, N.F. et al., 2012. Ultrafast Evolution and Loss of CRISPRs Following a Host Shift in a Novel Wildlife Pathogen, Mycoplasma gallisepticum. *PLoS Genetics*, 8(2), p.e1002511.

Dettman, J.R., Sztepanacz, J.L. & Kassen, R., 2016. The properties of spontaneous mutations in the opportunistic pathogen Pseudomonas aeruginosa. *BMC Genomics*, 17, p.27.

Didelot, X. et al., 2013. Genomic evolution and transmission of Helicobacter pylori in two South African families. *Proceedings of the National Academy of Sciences of the United States of America*, 110(34), pp.13880–5.

Didelot, X. et al., 2012. Microevolutionary analysis of Clostridium difficile genomes to investigate transmission. *Genome biology*, 13(12), p.R118.

Dillon, M.M. et al., 2016. Genome-wide biases in the rate and molecular spectrum of spontaneous mutations in Vibrio cholerae and Vibrio fischeri. *Molecular Biology and Evolution*, 34(1), pp.93–109.

Dillon, M.M. et al., 2015. The rate and molecular spectrum of spontaneous mutations in the GC-rich multi- chromosome genome of Burkholderia cenocepacia. *Genetics*, 200, pp.935–946.

Duchêne, S. et al., 2016. Genome-scale rates of evolutionary change in bacteria. *Microbial Genomics*, 2(October), pp.1-12.

Falush, D. et al., 2001. Recombination and mutation during long-term gastric colonization by Helicobacter pylori : Estimates of clock rates , recombination size , and minimal age. *Proceedings of the National Academy of Sciences of the United States of America*, 98(26), pp.15056–15061.

Feliziani, S. et al., 2014. Coexistence and Within-Host Evolution of Diversified Lineages of Hypermutable Pseudomonas aeruginosa in Long-term Cystic Fibrosis Infections. *PLoS Genetics*, 10(10), p.e1004651.

Feng, L. et al., 2008. A recalibrated molecular clock and independent origins for the cholera pandemic clones. *PloS one*, 3(12), p.e4053.

Ford, C.B. et al., 2011. Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection. *Nature Genetics*, 43(5), pp.482–486.

Grad, Y.H. et al., 2014. Genomic epidemiology of Neisseria gonorrhoeae with reduced susceptibility to cefixime in the USA: a retrospective observational study. *The Lancet Infectious diseases*, 14(3), pp.220–6. A

Gray, R.R. et al., 2011. Testing spatiotemporal hypothesis of bacterial evolution using methicillin-resistant staphylococcus aureus ST239 genome-wide data within a bayesian framework. *Molecular Biology and Evolution*, 28(5), pp.1593–1603.

Harris, S.R. et al., 2010. Evolution of MRSA During Hospital Transmission and Intercontinental Spread. *Science*, 327, pp.469–474.

Harris, S.R. et al., 2015. Genome specialization and decay of the strangles pathogen, Streptococcus equi, is driven by persistent infection. *Genome Research*, 25, pp.1360–1371.

Hawkey, J. et al., 2013. Evidence of microevolution of Salmonella Typhimurium during a series of egg-associated outbreaks linked to a single chicken farm. *BMC genomics*, 14(1), p.800.

He, M. et al., 2013. Emergence and global spread of epidemic healthcare-associated Clostridium difficile. *Nature genetics*, 45(1), pp.109–13.

Holden, M.T.G. et al., 2013. A genomic portrait of the emerfences, evolution, and global spread of methicillin-resistant Staphylococcus aureus. *Genome research*, 23, pp.653–664.

Holt, K. et al., 2016. Five decades of genome evolution in the globally distributed, extensively antibiotic resistant Acinetobacter baumannii global clone 1. *Microbial Genomics*, 2.

Holt, K.E. et al., 2012. Shigella sonnei genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nature Genetics*, 44(9), pp.1056–1059.

Howden, B. et al., 2013. Genomic Insights Howden, B., Holt, K., Lam, M., & Seemann, T. (2013). Genomic Insights to Control the Emergence of Vancomycin-Resistant Enterococci. *mBio*, 4(4), pp.e00412-13.

Kennemann, L. et al., 2011. Helicobacter pylori genome evolution during human infection. *Proceedings of the National Academy of Sciences of the United States of America*, 108(12), pp.5033–5038.

Klemm, E.J. et al., 2016. Emergence of host-adapted Salmonella Enteritidis through rapid evolution in an immunocompromised host. *Nature Microbiology*, 1, p.15023.

Kucukyildirim, S. et al., 2016. The Rate and Spectrum of Spontaneous Mutations in Mycobacterium smegmatis , a Bacterium Naturally Devoid of the Postreplicative Mismatch Repair Pathway. *G3*, 6, pp.2157–2163.

Lieberman, T.D. et al., 2011. Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nature Genetics*, 43(12), pp.1275–1280.

Lind, P. a & Andersson, D.I., 2008. Whole-genome mutational biases in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 105(46), pp.17878–17883.

Long, H. et al., 2015. Background mutational features of the radiation-resistant bacterium Deinococcus radiodurans. *Molecular biology and evolution*, 32(9), pp.2383–2392.

Long, H. et al., 2018. Evolutionary determinants of genome-wide nucleotide composition. *Nature Ecology and Evolution*, 2(February), pp.1–4.

Markussen, T. & Marvig, R.L., 2014. Environmental Heterogeneity Drives Within-Host Diversification and Evolution of Pseudomonas aeruginosa. *mBio*, 5(5), pp.e01592-14.

Marvig, R.L. et al., 2013. Genome analysis of a transmissible lineage of pseudomonas aeruginosa reveals pathoadaptive mutations and distinct evolutionary paths of hypermutators. *PLoS genetics*, 9(9), p.e1003741.

Mather, A.E., 2013. Distinguishable Epidemics of Multidrug-Resistant Salmonella Typhimurium DT104 in Different Hosts. *Science*, 341, pp.1514–1518.

Moran, N. a, McLaughlin, H.J. & Sorek, R., 2009. The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. *Science (New York, N.Y.)*, 323(5912), pp.379–382.

Morelli, G., Didelot, X., et al., 2010. Microevolution of Helicobacter pylori during prolonged infection of single hosts and within families. *PLoS Genetics*, 6(7), p.e1001036.

Morelli, G., Song, Y., et al., 2010. Yersinia pestis genome sequencing identifies patterns of global phylogenetic diversity. *Nature Genetics*, 42(12), pp.1140–1143.

Mutreja, A. et al., 2011. Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature*, 477, pp.462–465.

Nasser, W. et al., 2014. Evolutionary pathway to increased virulence and epidemic group A Streptococcus disease derived from 3,615 genome sequences. *Proceedings of the National Academy of Sciences*, 111(17), pp.E1768–E1776.

Njamkepo, E. et al., 2016. Global phylogeography and evolutionary history of Shigella dysenteriae type 1. *Nature Microbiology*, (March), p.16027.

Nübel, U. et al., 2010. A Timescale for Evolution, Population Expansion, and Spatial Spread of an Emerging Clone of Methicillin-Resistant Staphylococcus aureus. *PLoS Pathogens*, 6(4), p.e1000855.

Okoro, C.K. et al., 2012. Intracontinental spread of human invasive Salmonella Typhimurium pathovariants in sub-Saharan Africa. *Nature Genetics*, 44(11), pp.1215–1221.

Read, T.D. et al., 2013. Comparative analysis of Chlamydia psittaci genomes reveals the recent emergence of a pathogenic lineage with a broad host range. *mBio*, 4(2), pp.e00604-12.

Reeves, P.R. et al., 2011. Rates of Mutation and Host Transmission for an Escherichia coli Clone over 3 Years. *PloS One*, 6(10), p.e26907.

Roetzer, A. et al., 2013. Whole Genome Sequencing versus Traditional Genotyping for Investigation of a Mycobacterium tuberculosis Outbreak: A Longitudinal Molecular Epidemiological Study. *PLoS Medicine*, 10(2), p.e1001387.

Sánchez-Busó, L. et al., 2014. Recombination drives genome evolution in outbreak-related Legionella pneumophila isolates. *Nature Genetics*, 46(11), pp.1205–1211.

Schuenemann, V.J. et al., 2013. Genome-Wide Comparison of Medieval and Modern Mycobacterium leprae. *Science*, 341, pp.179–183.

Schultz, M.B. et al., 2016. Repeated local emergence of carbapenem resistant Acinetobacter baumannii in a single hospital ward. *Microbial Genomics*, 2(October), pp.1-15.

Senra, M.V.X. et al., 2018. An unbiased genome-wide view of the mutation rate and spectrum of the endosymbiotic bacterium Teredinibacter turnerae. *Genome Biology and Evolution*, 10(March), pp.723–730.

Smyth, D.S. et al., 2010. Population structure of a hybrid clonal group of methicillin-resistant Staphylococcus aureus, ST239-MRSA-III. *PLoS ONE*, 5(1), p.e8582.

Steglich, M. et al., 2015. Tracing the Spread of Clostridium difficile Ribotype 027 in

Germany Based on Bacterial Genome Sequences. *Plos One*, 10(10), p.e0139811.

Stinear, T.P. et al., 2014. Adaptive change inferred from genomic population analysis of the ST93 epidemic clone of community-associated methicillin-resistant Staphylococcus aureus. *Genome Biology and Evolution*, 6(2), pp.366–378.

Strauss, C. et al., 2017. Genome-wide mutation rate response to pH change in the coral reef. *mBio*, 8(4), pp.e01021-17.

Sun, Y. et al., 2017. Spontaneous mutations of a model heterotrophic marine bacterium. *ISME Journal*, 11(7), pp.1713–1718.

Sung, W. et al., 2015. Asymmetric Context-Dependent Mutation Patterns Revealed through Mutation – Accumulation Experiments Article Fast Track. *Molecular biology and evolution*, 32(7), pp.1672–1683.

Sung, W. et al., 2012. Drift-barrier hypothesis and mutation-rate evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 109, pp.18488–92.

Sung, W. et al., 2016. Evolution of the Insertion-Deletion Mutation Rate Across the Tree of Life. *G3*, 6, pp.2583–2591.

Tamas, I. et al., 2002. 50 Million Years of Genomic Stasis in Endosymbiotic Bacteria. *Science*, 296(5577), pp.2376–2379.

Uhlemann, A.-C. et al., 2014. Molecular tracing of the emergence, diversification, and transmission of S. aureus sequence type 8 in a New York community. *Proceedings of the National Academy of Sciences*, 111(18), pp.6738–6743.

Vandelannoote, K. et al., 2017. Multiple introductions and recent spread of the emerging human pathogen Mycobacterium ulcerans across Africa. *Genome Biology and Evolution*, 9(3), pp.414–426.

Walker, T.M. et al., 2013. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *The Lancet Infectious diseases*, 13(2), pp.137–146.

Ward, M.J. et al., 2014. Time-scaled evolutionary analysis of the transmission and antibiotic resistance dynamics of Staphylococcus aureus clonal complex 398. *Applied and Environmental Microbiology*, 80(23), pp.7275–7282.

Williamson, D.A. et al., 2017. Genomic insights into a sustained national outbreak of Yersinia pseudotuberculosis. *Genome Biology and Evolution*, 8(12), pp.3806–

3814.

Wilson, D.J. et al., 2009. Rapid evolution and the importance of recombination to the gastroenteric pathogen Campylobacter jejuni. *Molecular Biology and Evolution*, 26(2), pp.385–397.

Yang, L. et al., 2011. Evolutionary dynamics of bacteria in a human host environment. *Proceedings of the National Academy of Sciences*, 108(18), pp.7481–7486.

Young, B.C. et al., 2012. Evolutionary dynamics of Staphylococcus aureus during progression from carriage to disease. *Proceedings of the National Academy of Sciences of the United States of America*, 109(12), pp.4550–4555.

Zhou, Z. et al., 2013. Neutral Genomic Microevolution of a Recently Emerged Pathogen, Salmonella enterica Serovar Agona. *PLoS Genetics*, 9(4), p.e1003471.

Zhou, Z. et al., 2014. Transient Darwinian selection in Salmonella enterica serovar Paratyphi A during 450 years of global spread of enteric fever. *Proceedings of the National Academy of Sciences*, 111(33), pp.12199–12204.