Mutation rate dynamics reflect ecological change in an emerging zoonotic pathogen

Gemma G. R. Murray[1,*], Andrew J. Balmer[1], Josephine Herbert[1,2], Nazreen F. Hadijirin[1], Caroline L. Kemp[1], Marta Matuszewska[1], Sebastian Bruchmann[1], A. S. Md. Mukarram Hossain[1,3], Marcelo Gottschalk[4], A. W. (Dan) Tucker[1], Eric Miller[1,5,¶] and Lucy A. Weinert[1,¶]

1. Department of Veterinary Medicine, University of Cambridge, UK.
2. Current affiliation: School of Pharmacy and Biomedical Science, University of Portsmouth, UK.
3. Current affiliation: Cancer Biomarker Centre, Cancer Research UK Manchester Institute, The University of Manchester, Alderley Park, UK.
4. Département de Pathologie et Microbiologie, Université de Montréal, Canada.
5. Current affiliation: Haverford College, Pennsylvania, USA.

* Corresponding author: ggrmurray@gmail.com
¶ These authors are joint senior authors.

Short title: Mutation rate dynamics in an emerging zoonotic pathogen

Keywords: mutation rate, bacterial pathogenesis, genome size, *Streptococcus suis*

Abstract

While mutation is often deleterious, it can also be adaptive. Mutation rates vary both within and between bacterial species, and understanding what drives this variation is essential for understanding the evolutionary dynamics of bacterial populations. In this study we investigate two factors that are predicted to influence the trade-off between the costs and benefits of mutation: ecology and genome size. To investigate the relationship between these factors and mutation rate we conducted mutation accumulation experiments on eight strains of the emerging zoonotic pathogen *Streptococcus suis*. Natural variation within this species allows us to compare tonsil carriage and invasive disease isolates, from both more and less pathogenic populations, with a wide range of genome sizes. We find that invasive disease isolates have repeatedly evolved mutation rates that are higher than those of closely related carriage isolates, regardless of variation in genome size. Independent of this variation in overall rate, we also observe a stronger bias towards G/C to A/T mutations in isolates from more pathogenic populations, whose genomes tend to be smaller and more AT-rich. Our results suggest that ecology is a stronger correlate of mutation rate than genome size over these timescales, and that transitions to invasive disease are consistently accompanied by rapid increases in mutation rate. These results shed light on the impact ecology can have on the adaptive potential of bacterial pathogens.

Introduction

Mutation rates vary within and between bacterial species, contributing to differences in both the burden of deleterious mutations and the capacity to adapt to environmental change [1–3]. Understanding how mutation rates evolve in response to selective pressures is fundamental to our understanding of evolutionary dynamics. In the case of bacterial pathogens, it is important for understanding pathogen emergence, evasion of host immunity, and evolution of antimicrobial resistance [4–7].

There are several reasons to predict that bacterial pathogens will have higher mutation rates than non-pathogens. First, pathogens may be less able to constrain the evolution of their mutation rate. Mutation rates evolve as linked traits, and as deleterious mutations are more common that beneficial mutations, theory predicts that selection against these mutations will generally lead to indirect selection for lower mutation rates. Common features of pathogen ecologies such as host-restriction, frequent between-host transmission, and rapid within-host adaptation, could all contribute to a smaller effective population size [4,8]. This could lead to a reduced efficacy of natural selection, and therefore to maladaptive evolution, including a higher mutation rate [1,9,10]. Second, a higher mutation rate may be associated with a lower selective cost for pathogens. Bacterial pathogens tend to have smaller genomes with fewer genes than closely related non-pathogens [13,14], and a smaller genome will mean that a higher (per site) mutation rate will lead to fewer additional mutations [15]. Finally, pathogens might actually benefit from higher mutation rates. Pathogens often face challenging and hostile environments, and higher mutation rates can increase the speed of adaptation [3,11,12]. This might lead to positive selection for a higher mutation rate, although indirectly via linkage with beneficial mutations.

Despite these predictions, the influence of ecology and genome size on mutation rate dynamics in natural populations of bacterial pathogens is not well understood. Across bacterial species, mutation rates have been found to be inversely correlated with genome size, and largely independent of ecology [1,16]. Within several bacterial species, hypermutable strains with orders of magnitude higher mutation rates than the species average have been identified [17–23]. Nevertheless, this within-species variation remains poorly characterized, and while there is evidence that transient increases in mutation rates can promote adaptation, it remains unclear whether there is an association with pathogenic ecologies.

To investigate the influence of ecology and genome size on mutation rate, we carried out mutation accumulation (MA) experiments and whole-genome sequencing for eight isolates of the opportunistic and emerging bacterial pathogen *Streptococcus suis*. This experimental approach allowed us to obtain precise estimates of mutation rate, and therefore to investigate small-scale rate variation, which is more likely to persist over time, and to be associated longer-term ecological transitions [24,25]. *S. suis* both asymptomatically colonises the upper respiratory tract of pigs, and causes severe invasive infections in both pigs and humans [26,27]. Natural variation within this species allows us to compare the mutation rates of closely related isolates with different pathogenic ecologies and a range of genome sizes (1.97 – 2.67 Mb) [14,28].

Results

We first estimated the mutation rates of two pairs of closely related isolates, that span the known range of *S. suis* genome size (strains 1-4 in Figures 1 and S1, and Tables 1 and S1). In each pair, one isolate was sampled from the site of an invasive infection (disease) and the other from the tonsils of a pig without *S. suis* associated disease (carriage). The relationship between these isolates, and their placement in a core genome phylogeny of *S. suis* was established in a previous study [14]. This study found that while both asymptomatic carriage isolates and invasive disease isolates are present across the *S. suis* phylogeny, invasive disease isolates are more common in one clade (a “more pathogenic” group) and carriage isolates more common outside of this clade (a “less pathogenic” group). Genetic diversity in the more pathogenic group is lower than, and falls within the diversity of the less pathogenic group, consistent with a more recent origin [11]. One of our pairs of isolates was sampled from this more pathogenic group (isolates 1 and 2) and the other from a less pathogenic group (isolates 3 and 4) (Figure 1a and S1). Our choice of strains therefore allows us to discriminate between three possible correlates of mutation rate: short-term transitions from carriage to disease (Figure 1b), long-term changes in pathogenicity (Figure 1c), and genome size (Figure 1d).

We estimated the mutation rates of the four isolates through four parallel 200-day MA experiments. We evolved 75 replicate lines of each strain over 200 days, with daily passaging through single-colony bottlenecks. We estimated that this 200-day period corresponds to between 3,445 to 3,991 generations for these four strains (Table 1 and S2, and Figure S2). We then sequenced the genomes of 50 randomly selected evolved lines of each strain, and estimated mutation rates through identifying differences in the genomes of the evolved lines compared to the genomes of the ancestral strains.

Increased mutation rates associated with the transition from carriage to disease

We found that the two disease isolates had accumulated single-base mutations at a faster rate (per site per generation) than the two carriage isolates, while there was no consistent difference in overall rate between isolates from the more and less pathogenic groups, or a correlation with genome size (Figure 1, Tables 1 and S3). The accumulation rate in disease isolates is observed across all classes of single-base mutations, including transitions, transversions, A/T to G/C transitions and transversions, and G/C to A/T transitions and transversions (Figures 2 and S3), and across both core and accessory genes (Figure S4).

MA experiments aim to minimise the effect of selection on new mutations, so that accumulation rates reflect mutation rates as closely as possible. Our results indicate that the influence of selection on accumulation rates was low across our four MA experiments. First, distributions of single-base mutations across the 50 replicate lines do not differ significantly from a Poisson distribution for any of the four strains (Table S4), suggesting that accumulation rates did not vary across lines. Second, rates of single-base mutation were similar across 1st and 2nd codon positions and 4-fold degenerate sites for each strain (Figure S5). Third, non-unique single base mutations were rare: only 5 of the 2,267 single base changes observed over the four experiments occurred in more than one line, and none occurred in more than two. Fourth, we found no evidence of an overall change in maximum growth rates over the course of the experiment for the evolved lines of the two carriage strains or the disease strain with the smaller genome, and a net decline in maximum growth rate in the disease strain with the larger genome, consistent with its accumulation of the largest number of mutations, and these mutations tending to have a weakly deleterious effect (Figures S6, S7, Table S5). And finally, we sequenced five lines of each of the four strains at the midpoint of the experiment (100 days), and estimated rates of accumulation over the first and second halves of the experiment. This revealed no evidence of a change in rate (Table S6).

To further investigate the observed difference in mutation rate between disease and carriage isolates we undertook an additional smaller scale (25-day) MA experiment with an additional four isolates. These four isolates were chosen to include much closer relatives of the two isolates from the more pathogenic group from our original experiment, and a more distantly related disease/carriage pair, also from the more pathogenic group (Tables 1, S1 and S7). This allowed us to investigate whether the difference between disease and carriage isolates holds over shorter evolutionary distances, and to test the generality of the association. 15 replicate lines of the four

strains were evolved over 25 days, 11-13 randomly selected lines of each strain were sequenced, and mutations were identified in the same way as in the 200-day experiment.

Our estimates of single-base mutation rates for this combined data set of four pairs of disease and carriage isolates showed a consistent pattern: disease isolates have faster rates than closely related carriage isolates. Moreover, all disease isolates have faster rates than all carriage isolates. While there is uncertainty associated with the rate estimates for individual strains and our sample size is small, a two-sided paired t-test suggests that the difference between disease and carriage isolates we observe in our point estimates is unlikely to have arisen by chance: $p = 0.035$, Figure 3, Table S8). This pattern appears to be independent of genome size as it holds across pairs with variable average genome sizes, and both across pairs in which the disease strain has a smaller genome than the carriage isolate (3 pairs) and a pair in which the disease isolate has a larger genome than the carriage isolate (Figure 3b,c). While we observe greater variation in overall mutation rates across the four carriage isolates in our two experiments than across the four disease isolates, this variation is not correlated with genome size (Figures 3 and S8).

Changes in mutational spectrum associated with increased pathogenicity

Over the course of our 200-day MA experiment sufficient mutations accumulated to allow us to examine differences in mutational bias across isolates (Table S3). Comparing the four isolates from the 200-day experiment, we found that while there is no consistent difference in the single-base mutation rate across isolates from more and less pathogenic groups there is a difference in the mutational spectrum. The two isolates from the more pathogenic group have a higher rate of transversions relative to transitions, and a higher rate of G/C to A/T transitions relative to A/T to G/C transitions (Figure 1f,g). In addition, while we observe a context-dependency of mutation rates (a higher mutation rate at sites flanked by a G/C base) in the two isolates from the less pathogenic group as in previous studies [29,30], we do not observe this in the two isolates from the more pathogenic group (Figure S9).

The differences in mutational bias across the different groups are observed consistently across both core and accessory genes (Figure S10). Moreover, the rate of G/C to A/T mutations relative A/T to G/C mutations is correlated with the base composition of core genes (despite a small sample size, Pearson's correlation: $r^2 = 0.98$, $p = 0.01$; Figure 4). Comparison of our estimates of mutational bias and the base composition of core genes across these four strains also reveals that the two isolates from the more pathogenic group are further from their equilibrium base composition (Figure S11).

This suggests that the change in bias occurred in an ancestor of the two isolates from the pathogenic group, and may therefore be associated with their transition to a more pathogenic ecology.

Deletion rates decline with genome reduction

Deletion mutations show a different pattern of variation to single-base mutations across the four strains in the 200-day MA experiment. When considering short insertions and deletions (< 30 bp), we observe a faster rate of deletion in the two isolates from the less pathogenic group, without a corresponding increase in the rate of insertion (Figure 2g, Figure S12, Table S9). As most small deletions occur in intergenic regions, which tend to be shorter in the two isolates from the more pathogenic group, the difference may be due to the absence of intergenic regions that are more prone to deletion in these isolates.

In contrast, larger deletions (> 100 bp) arose at a faster rate in the two carriage isolates than the two disease isolates (particularly the carriage isolate from the less pathogenic group, Figure 2h). We identified 34 deletions larger than 100 bp across the four strains, with 10 larger than 10 kb (Table S10). Most involved the loss of regions that show signatures of being mobile genetic elements (including phage-associated genes), which are more common in the isolates with larger genomes (Figure S13, Table S11, S12).

No loss of genes from DNA-repair pathways

We found no evidence that the loss or truncation of genes from known DNA repair pathways led to the observed differences in mutation rate or spectrum. Genes from the base excision repair, nucleotide excision repair and mismatch repair pathways were uniformly present across all eight strains (Table S14). In addition, no genes that were consistently absent in disease strains were also consistently present in carriage strains (or vice versa) (Tables S13). Some DNA-repair pathway genes were present in multiple copies in some strains, but the presence/absence of these additional copies did not correlate with disease/carriage status or location in more/less pathogenic groups. In addition, the majority of DNA-repair pathway genes had a constant length across the eight strains, and where small length variations were observed, these did not correlate with disease/carriage status or presence in a more/less pathogenic group (Table S14).

Discussion

While we were only able to estimate the mutation rates of a small sample of *S. suis* isolates due to the time-consuming and labour-intensive nature of MA experiments, our results indicate two possible links between pathogenic ecologies and the rate and spectrum of mutations. First, *S. suis* isolates sampled from the site of invasive infections have consistently higher mutation rates than carriage isolates sampled from the tonsils. Second, isolates from the more pathogenic group of *S. suis* show a greater bias towards both G/C to A/T mutations and transversions, than those from a less pathogenic group. These changes in bias are uncoupled from disease/carriage status, and from overall mutation rate. In contrast, while genome size has previously been found to be a strong predictor of mutation rate variation across bacterial species (Figure S8) [1,16], we found no evidence of a link with mutation rates within *S. suis*. Disease isolates with a wide range of genome sizes have consistent mutation rates, and while the mutation rates of carriage isolates are more variable, they are not correlated with genome size. Nevertheless, as our sample size is small, further work is required to definitively test these associations.

In opportunistic pathogens, such as *S. suis*, carriage is thought to be a prerequisite of invasive disease, with carriage isolates acting as a source of disease isolates [31]. This transition from asymptomatic carriage to invasive disease is likely to involve both growth in novel environments, and additional pressures from host immune systems [32]. Under these conditions, isolates with faster mutation rates may have a selective advantage because they are more likely to generate novel adaptive variants [2,3,22], or because mutation rate is linked to another trait that is adaptive in these new conditions [e.g. 33,34]. While the strains in our experiments only represent a single isolate from any host, the difference we observe between carriage and disease isolates plausibly reflects within-host evolution. If this is the case, the rapidity of the change in rates suggests that it is driven by positive selection rather than random drift. This might involve selection on standing variation in carriage populations; variants with higher mutation rates could be present at low frequencies in carriage populations and rise to high frequencies in invasive populations. This is consistent with the results of previous studies that have identified variation in mutation rates in within-host bacterial populations [6,18,21,23,35], and with our observation of no correlation between the difference in rate and the evolutionary distance between disease and carriage isolates.

In contrast to previous studies that have identified hypermutable strains, the rate variation we observe between disease and carriage isolates is not associated with the loss or pseudogenisation of genes in DNA-repair pathways, and is much smaller in scale. While this means a smaller increase

in the frequency of adaptive mutations, it also means a smaller increase in the frequency of deleterious mutations. These higher rates may therefore be maintained for longer and reach higher frequencies in a population than larger rate increases. This could contribute to the maintenance of mutation rate heterogeneity within populations, which is predicted to promote faster responses to selective challenges, while reducing the long-term burden of deleterious mutations [7,36].

In addition to the difference in point mutation rates across disease and carriage isolates, we observed a difference in deletion rates. While we only have the power to estimate rates of deletion in our 200-day MA experiment, almost all of the large deletion events in this experiment occurred in the two carriage isolates, and often involved the loss of regions that contain phage-associated genes. This difference in rates might reflect fewer temperate prophages in the genomes of disease isolates. In *S. suis* the transition from carriage to disease is associated with a reduction in genome size, and it has been suggested that this is, in part, due to the loss of mobile genetic elements [14]. While the rate of prophage loss observed in our experiments is probably too slow to explain genome reduction during the transition from carriage to disease within a host, the stress associated with this transition might lead to a higher rate of loss [37].

Previous studies have found evidence that several clusters of *S. suis* have evolved to become more pathogenic to pigs, and are also responsible for zoonotic disease in humans [14,28]. The emergence of these more pathogenic clusters has previously been shown to be associated with both genome reduction and an increased AT-richness of the core genome [14]. Here, we have found evidence that it is also associated with a change in mutational spectrum. While this is only supported by comparison of the four strains from our 200-day experiment (due to too few mutations in the 25-day MA experiment) our observation of a strong correlation between mutational bias and core genome base composition, is consistent with this underlying a broader association between AT-richness and pathogenicity in *S. suis* that was identified in a previous study [14]. Increased AT-richness is a common correlate of genome reduction in bacterial symbionts [8,9,38], and has also been observed in the evolution of the pathogen *Shigella* [39]. As G/C to A/T mutations tend to be more deleterious than A/T to G/C mutations, these patterns are commonly attributed to increased genetic drift [8,9]. While our results are consistent with this as an ultimate explanation of the increased AT-richness in more pathogenic clusters of *S. suis*, the correlation we observe between mutational bias and core genome nucleotide composition suggests that it is an immediate consequence of a change in mutational bias. In addition, our observation that the change in bias is not accompanied by an

increase in overall rate suggests that it is unlikely to lead to a substantial increase in the burden of deleterious mutations [40,41].

Overall, our results indicate that mutation rate variation within bacterial species extends beyond hypermutable strains, and can be sensitive to ecological transitions. Both the changes in overall rate and in mutational bias we observe in *S. suis* could influence the frequency of adaptive mutations, and therefore the capacity to rapidly respond to selective challenges such as evading host immunity or antibiotic treatments [42,43]. This link between ecology and mutation rate could therefore prove to be both important in understanding the evolutionary trajectories of bacterial pathogens, and a useful marker of ecological change.

Methods

Strains and culture conditions

Strains were selected from a collection of *S. suis* isolates described in [14] together with three isolates sampled from pig farms in Europe (Table S2). Isolates were sampled from different pig farms in Canada, Spain, the Netherlands and the UK, between 2009 to 2017. Isolates were classified as "disease" if they were recovered from systemic sites in pigs or humans with clinical signs consistent with *S. suis* infection, or from the lungs of a pig with signs of pneumonia. Isolates recovered from the tonsils or tracheo-bronchus of healthy pigs or pigs without signs of *S. suis* infection were classified as "carriage". Strains for both experiments were chosen based on clinical information, genome size, and location in a core genome phylogeny previously described in [14]. All strains were from pigs, and all "disease" isolates were associated with systemic infections.

The solid media used for all experiments was Todd-Hewitt broth (THB) supplemented with 1.5% (w/v) agarose and 0.2% yeast extract (THY) (Oxoid, Basingstoke). 20% Glycerol/THB (36.4g in 1 litre) with 0.2% yeast extract was used for archiving and PCR, and THB with 0.2% yeast extract was used for overnight growth and growth rate experiments. Cultures were streaked to single colonies on solid media and incubated overnight in a static incubator at 37°C to generate stock plates. For overnight cultures, an independent colony was picked from the stock plate to inoculate THB + 0.2% yeast extract broth and incubated in a static incubator at 37°C.

Mutation accumulation experiments

75 replicate lines were established from each strain in the 200-day mutation accumulation (MA) experiment and 15 replicate lines in the 25-day MA experiment. Each line was passaged through single-colony bottlenecks by selecting the last visible independent colony in the streak (to minimise selection bias). Plates were incubated at 37°C for approximately 24 hours between each transfer. All passaged plates were stored at 4°C for 24 hours, to allow lines that failed to grow during passage to be reset from the stored plate. In the 200-day experiment lines were archived in liquid broth every 14 days, and every 100 days.

PCR

For the 200-day experiment, the four strains were grown on quadrants of a single plate, and to allow for errors during passaging to be corrected we developed a multiplex PCR to distinguish the strains. Two sets of primer sequences for a multiplex PCR were designed using a custom Python script to identify unique regions of each strain of set lengths that would resolve in electrophoresis, along with

a *S. suis* positive control (Table S15). To conduct the PCR, a single colony was transferred to 150 μl liquid media, 5μl of which was added to 50μl sterile MiliQ water and heated at 95°C for six minutes. Amplification conditions and reagent volumes are given in Table S15. We tested all lines every 14 days. If a mismatch was identified, lines were set back to the previous PCR run.

Estimation of generation times

Strains were streaked to single colonies from overnight cultures on THY +0.2% yeast extract agar plates and incubated at 37°C for 24 hours, with a minimum of three biological replicates. Single colonies were collected by excising a small disc of agar around a colony and resuspending it in 10ml of PBS. The solution was serially diluted, and dilutions spread on THY +0.2% yeast extract agar plates. The plates were incubated in a static incubator at 37°C and examined after 24 hours to determine the number of colony forming units (CFU). Generation times were calculated by dividing log2(average CFU) by the period of growth (Table S2, Figure S2).

Growth rate measurements

Growth rates were estimated for the ancestral lines and a random sample of 25 of the sequenced evolved lines of each strain in the 200-day experiment, with a minimum of three biological replicates. Overnight cultures were pelleted by centrifugation at 4000xg for 3 minutes to remove spent media. After discarding the supernatant, the cell pellet was resuspended in fresh THY +0.2% yeast extract media to a final concentration of 10^7 CFU per well. 300µl of the culture was transferred into wells and incubated in a Bioscreen C (Oy Growth Curves Ab Ltd) at 37°C, with optical density (OD_{600}) measured every 5 minutes for 24 hours. For 10 evolved lines (9 of strain 1 and one of strain 3) there was insufficient overnight growth to reach the required starting concentration, and growth rates could not be accurately measured. Maximum growth rates were calculated by taking the slope of a linear regression model of $\log_2(OD_{600})$ over time, using a 30-minute sliding window to identify the period of fastest growth (for details, see [44]). To test the reliability of OD as a proxy for CFU during exponential growth, additional time-sampled growth curves were completed, measuring both OD and CFU for each of the four ancestral strains (Table S16).

Sequencing

Illumina whole genome sequencing was undertaken for all ancestral strains, a random sample of 50 evolved lines of each of the strains in the 200-day MA experiments, 5 evolved lines of each of the strains in the 200-day MA experiments at day 100, and 11-13 evolved lines of each of the strains in

the 25-day MA experiments. DNA extraction, library preparation and sequencing using a HiSeq 2500 instrument (Illumina, San Diego, CA, USA) was undertaken by MicrobesNG (Birmingham, UK).

Long-read sequencing, assembly and annotation of ancestral strains

We assembled high-quality reference genomes for all eight ancestral strains using methods that combine short-read and long-read sequence data. For the 200-day experiments, long-read sequencing library preparation was performed using Genomic-tips and a Genomic Blood and Cell Culture DNA Midi kit (Qiagen, Hilden, Germany). Sequencing was performed on the Sequel instrument from Pacific Biosciences using v2.1 chemistry and a multiplexed sample preparation. Reads were demultiplexed using Lima in the SMRT link software (https://github.com/PacificBiosciences/barcoding). Reads shorter than 2500 bases were removed using prinseq-lite.pl (https://sourceforge.net/projects/prinseq/). Hybrid assemblies, using filtered PacBio and Illumina reads, and preliminary assemblies of long-read data generated with Canu v1.9 [45], were generated with Unicycler v0.4.7 using the normal mode and default settings [46]. Assembly graphs were visualised and, if necessary, manually corrected with Bandage [47]. For the 25-day MA experiments, library preparation, short-read and long-read sequencing, and hybrid assembly with Unicycler [46] were undertaken by MicrobesNG as part of their Enhanced Genome Service, which uses both Illumina and Oxford Nanopore Technologies. For 4/8 strains the assemblies were single-contig, and for the other four the longest contig was >98% of the total assembly length (smaller contigs likely representing plasmids or mobile elements).

Genomes were annotated using Prokka v1.14.5 [48] and putative mobile elements identified using IslandViewer 4 [49]. Panaroo v1.2.2 [50] was used to identify orthologous genes (using recommended parameter settings) and create alignments of shared genes. Core-genome distance matrices were estimated and a neighbour-joining tree created using these alignments and the *ape* package in *R* [51].

Mutation calling in evolved lines

Mutations in evolved lines were called by mapping short-read sequence data to the reference genomes of the ancestral strains. Illumina reads were adapter-trimmed using Trimmomatic with a sliding window quality cutoff of Q15 [52]. They were mapped to the ancestral strain (excluding any short contigs) using Bowtie2, and variants called using SAMtools and BCFtools [53,54]. False variant calls were identified using several approaches. First, we excluded any calls with a depth of less than six reads (average coverage was always >30x) or where the reference allele was present

in more than 5% of reads. Second, short-read data from the ancestral strains was mapped back to itself, and any variants identified were excluded. Finally, any variant calls that either had a lower quality score than the maximum for the line, were within 100 bases of another variant call in any line, or were within 1000 bases of the start or end of the reference genome, were checked by eye for evidence of mapping error. In the 200-day experiment, clusters of mutations were identified (mutations within 30 bases of another in the same line). They were common only in strain 4 (<5% of mutations in the other three strains), and only 10% fell in regions of the genome shared across the 4 strains. These mutations were not excluded from our core analyses, but Figures S4 and S5 describe the impact of their exclusion.

Indels were identified using both SAMtools and ScanIndel [54,55]. To avoid false calls, we required indels to be identified by both analyses and not identified when short-read data for the ancestral strain was mapped back to the reference assembly. In addition, we required that the alternate allele is supported by at least five reads, at least one in each orientation, and fewer than 5% of reads supporting the reference allele, a quality score in SAMtools of at least 20, and an 'IMF' of at least 0.8. Deletions longer than 100 bases were called by identifying extended regions with zero coverage in our mapped assemblies, and confirmed through examination by eye.

Estimating average rates and statistical comparisons

The single-base mutation rate per site per generation for each strain was estimated as:

$$\mu_{bs} = \frac{m}{ng}$$

where m is the number of observed single-base mutations, n is the number of nucleotide sites analysed (genome length multiplied by the number of lines), and g is the total number of generations over the duration of the experiment (see above). This value was estimated for the whole genome, for different subsets of sites, and for different types of single base mutations, with the number of nucleotide sites analysed adjusted in each case. 95% confidence intervals were generated for estimates using two approaches. First, based on 1000 bootstrap samples of the lines of each strain, and second based on estimates of the standard deviation in the rate across lines. We found that confidence intervals were similar across the two methods, and none of our results were contingent on the use of either method.

Identification of DNA-repair genes

DNA-repair genes were identified from the Prokka and Panaroo annotations using descriptions of genes in the *S. suis* DNA mismatch repair, base excision repair and nucleotide excision repair pathways from KEGG [56].

Data Availability

All sequence data from the two MA experiments have been deposited on the (lines short read archive under the BioProject ID: PRJNA763153. Assemblies of the eight ancestral strains have been deposited under the BioProject ID: PRJNA763404.

Acknowledgements

This work was primarily funded by an Isaac Newton Trust / Wellcome Trust ISSF / University of Cambridge Joint Research Grant to LAW. In addition, LAW, ELM, GGRM were supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (109385/Z/15/Z). GGRM was also supported by a ZELS BBSRC award (BB/L018934/1) and a Research Fellowship at Newnham College. AJB was supported by a BBSRC-funded studentship. MM was funded by the Medical Research Council, co-funded by the Raymond and Beverly Sackler Fund. MH was supported by EU Horizon 2020 grant “PIGSs”, which also funded the collection of additional isolates in this study. Genome sequencing was undertaken by MicrobesNG (http://www.microbesng.uk/), which is supported by the BBSRC (grant number BB/L024209/1), and by the Sanger Institute in collaboration with Julian Parkhill. Finally, we wish to thank John Welch and Olivier Restif for helpful comments on the manuscript, and Virginia Aragon and her team for one of our strains.

Author Contributions

LAW, ELM and NFH designed the 200-day MA experiment. ELM, NFH, AJB and CLK conducted the 200-day MA experiments, with assistance from MM, MH and LAW. LAW, GGRM and JH designed the 25-day MA experiments. JH conducted the 25-day MA experiments. AJB conducted the growth curves. SB assembled reference genomes for the 200-day MA experiments. GGRM designed the mutation-calling pipeline for both experiments, and conducted the bioinformatic analysis for the 200-day MA experiments. JH conducted the bioinformatic analysis for the 25-day MA experiments. GGRM wrote the paper with contributions from all other authors.

Declaration of interests

The authors declare no competing interests.

References

1. Lynch M, Ackerman MS, Gout JF, Long H, Sung W, Thomas WK, et al. Genetic drift, selection and the evolution of the mutation rate. Nature Reviews Genetics. Nature Publishing Group; 2016. pp. 704–714. doi:10.1038/nrg.2016.104
2. Denamur E, Matic I. Evolution of mutation rates in bacteria. Molecular Microbiology. John Wiley & Sons, Ltd; 2006. pp. 820–827. doi:10.1111/j.1365-2958.2006.05150.x
3. Giraud A, Radman M, Matic I, Taddei F. The rise and fall of mutator bacteria. Current Opinion in Microbiology. 2001;4: 582–585. doi:10.1016/S1369-5274(00)00254-X
4. Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. Within-host evolution of bacterial pathogens. Nature Reviews Microbiology. 2016;14: 150–162. doi:10.1038/nrmicro.2015.13
5. Bonneaud C, Weinert LA, Kuijper B. Understanding the emergence of bacterial pathogens in novel hosts. Philosophical Transactions of the Royal Society B: Biological Sciences. 2019;374: 20180328. doi:10.1098/rstb.2018.0328
6. Oliver A, Cantón R, Campo P, Baquero F, Blázquez J. High frequency of hypermutable Pseudomonas aeruginosa in cystic fibrosis lung infection. Science. 2000;288: 1251–1253. doi:10.1126/science.288.5469.1251
7. Matic I. Mutation Rate Heterogeneity Increases Odds of Survival in Unpredictable Environments. Molecular Cell. Cell Press; 2019. pp. 421–425. doi:10.1016/j.molcel.2019.06.029
8. Moran NA. Microbial minimalism: Genome reduction in bacterial pathogens. Cell. Cell Press; 2002. pp. 583–586. doi:10.1016/S0092-8674(02)00665-7
9. Moran NA, McLaughlin HJ, Sorek R. The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. Science. 2009;323: 379–382. doi:10.1126/science.1167140
10. Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M. Drift-barrier hypothesis and mutation-rate evolution. Proceedings of the National Academy of Sciences of the United States of America. 2012;109: 18488–18492. doi:10.1073/pnas.1216223109
11. Wiser MJ, Ribeck N, Lenski RE. Long-term dynamics of adaptation in asexual populations. Science. 2013;342: 1364–1367. doi:10.1126/science.1243357
12. Taddei F, Radman M, Maynard-Smith J, Toupance B, Gouyon PH, Godelle B. Role of mutator alleles in adaptive evolution. Nature. 1997;387: 700–702. doi:10.1038/42696
13. Weinert LA, Welch JJ. Why might bacterial pathogens have small genomes? Trends in Ecology and Evolution. Elsevier Ltd; 2017. pp. 936–947. doi:10.1016/j.tree.2017.09.006
14. Murray GGR, Charlesworth J, Miller EL, Casey MJ, Lloyd CT, Gottschalk M, et al. Genome reduction is associated with bacterial pathogenicity across different scales of temporal and ecological divergence. Molecular Biology and Evolution. 2021;38: 1570–1579. doi:10.1093/molbev/msaa323
15. Smith DR. The mutational hazard hypothesis of organelle genome evolution: 10 years on. Molecular Ecology. 2016;25: 3769–3775. doi:10.1111/mec.13742
16. Drake JW. A constant rate of spontaneous mutation in DNA-based microbes. Proceedings of the National Academy of Sciences. 1991;88: 7160–7164. doi:10.1073/pnas.88.16.7160
17. Painter PR. Mutator genes and selection for the mutation rate in bacteria. Genetics. 1975;79: 649–660.
18. LeClerc JE, Li B, Payne WL, Cebula TA. High mutation frequencies among Escherichia coli and Salmonella pathogens. Science. 1996;274: 1208–1211. doi:10.1126/science.274.5290.1208

19. Wang H, Xing X, Wang J, Pang B, Liu M, Larios-Valencia J, et al. Hypermutation-induced in vivo oxidative stress resistance enhances Vibrio cholerae host adaptation. Satchell KJF, editor. PLOS Pathogens. 2018;14: e1007413. doi:10.1371/journal.ppat.1007413
20. Matic I, Radman M, Taddei F, Picard B, Doit C, Bingen E, et al. Highly variable mutation rates in commensal and pathogenic Escherichia coli. Science. 1997;277: 1833–1834. doi:10.1126/science.277.5333.1833
21. Deiham B, Douraghi M, Adibhesami H, Yaseri M, Rahbar M. Screening of mutator phenotype in clinical strains of Acinetobacter baumannii. Microbial Pathogenesis. 2017;104: 175–179. doi:10.1016/j.micpath.2017.01.033
22. Swings T, van den Bergh B, Wuyts S, Oeyen E, Voordeckers K, Verstrepen KJ, et al. Adaptive tuning of mutation rates allows fast response to lethal stress in escherichia coli. eLife. 2017;6. doi:10.7554/eLife.22939
23. Ramiro RS, Durão P, Bank C, Gordo I. Low mutational load and high mutation rate variation in gut commensal bacteria. PLoS Biology. 2020;18: e3000617. doi:10.1371/journal.pbio.3000617
24. Halligan DL, Keightley PD. Spontaneous mutation accumulation studies in evolutionary genetics. Annual Review of Ecology, Evolution, and Systematics. 2009;40: 151–172. doi:10.1146/annurev.ecolsys.39.110707.173437
25. Barrick JE, Lenski RE. Genome dynamics during experimental evolution. Nature Reviews Genetics. Nature Publishing Group; 2013. pp. 827–839. doi:10.1038/nrg3564
26. Vötsch D, Willenborg M, Weldearegay YB, Valentin-Weigand P. *Streptococcus suis* - The "two faces" of a pathobiont in the porcine respiratory tract. Frontiers in Microbiology. Frontiers Media S.A.; 2018. doi:10.3389/fmicb.2018.00480
27. Lun ZR, Wang QP, Chen XG, Li AX, Zhu XQ. *Streptococcus suis*: an emerging zoonotic pathogen. Lancet Infectious Diseases. Elsevier; 2007. pp. 201–209. doi:10.1016/S1473-3099(07)70001-4
28. Weinert LA, Chaudhuri RR, Wang J, Peters SE, Corander J, Jombart T, et al. Genomic signatures of human and animal disease in the zoonotic pathogen *Streptococcus suis*. Nature Communications. 2015;6: 6740. doi:10.1038/ncomms7740
29. Blake RD, Hess ST, Nicholson-Tuell J. The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. Journal of Molecular Evolution. 1992;34: 189–200. doi:10.1007/BF00162968
30. Lee H, Popodi E, Tang H, Foster PL. 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 of the United States of America. 2012;109: E2774–E2783. doi:10.1073/pnas.1210309109
31. Weiser JN, Ferreira DM, Paton JC. Streptococcus pneumoniae: transmission, colonization and invasion. Nature Reviews Microbiology 2018 16:6. 2018;16: 355–367. doi:10.1038/s41579-018-0001-8
32. Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. Within-host evolution of bacterial pathogens. Nature Reviews Microbiology. Nature Publishing Group; 2016. pp. 150–162. doi:10.1038/nrmicro.2015.13
33. Torres-Barceló C, Cabot G, Oliver A, Buckling A, MacLean RC. A trade-off between oxidative stress resistance and DNA repair plays a role in the evolution of elevated

mutation rates in bacteria. Proceedings of the Royal Society B: Biological Sciences. 2013;280: 20130007. doi:10.1098/rspb.2013.0007
34. Richardson AR, Stojiljkovic I. Mismatch repair and the regulation of phase variation in Neisseria meningitidis. Molecular Microbiology. 2001;40: 645–655. doi:10.1046/j.1365-2958.2001.02408.x
35. Couce A, Alonso-Rodriguez N, Costas C, Oliver A, Blázquez J. Intrapopulation variability in mutator prevalence among urinary tract infection isolates of *Escherichia coli*. Clinical Microbiology and Infection. 2016;22: 566.e1-566.e7. doi:10.1016/j.cmi.2016.03.008
36. Alexander HK, Mayer SI, Bonhoeffer S. Population Heterogeneity in Mutation Rate Increases the Frequency of Higher-Order Mutants and Reduces Long-Term Mutational Load. Molecular biology and evolution. 2017;34: 419–436. doi:10.1093/molbev/msw244
37. Howard-Varona C, Hargreaves KR, Abedon ST, Sullivan MB. Lysogeny in nature: Mechanisms, impact and ecology of temperate phages. ISME Journal. Nature Publishing Group; 2017. pp. 1511–1520. doi:10.1038/ismej.2017.16
38. Ochman H, Moran NA. Genes lost and genes found: Evolution of bacterial pathogenesis and symbiosis. Science. 2001;292: 1096–1099. doi:10.1126/science.1058543
39. Balbi KJ, Rocha EPC, Feil EJ. The temporal dynamics of slightly deleterious mutations in *Escherichia coli* and *Shigella* spp. Molecular Biology and Evolution. 2009;26: 345–355. doi:10.1093/molbev/msn252
40. Haldane JBS. The effect of variation of fitness. The American Naturalist. 1937;71: 337–349. doi:10.1086/280722
41. Muller HJ. Our load of mutations. American journal of human genetics. 1950;2: 111–176. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1716299/
42. Payne JL, Menardo F, Trauner A, Borrell S, Gygli SM, Loiseau C, et al. Transition bias influences the evolution of antibiotic resistance in *Mycobacterium tuberculosis*. PLoS Biology. 2019;17: e3000265. doi:10.1371/journal.pbio.3000265
43. Stoltzfus A, McCandlish DM. Mutational biases influence parallel adaptation. Molecular Biology and Evolution. 2017;34: 2163–2172. doi:10.1093/molbev/msx180
44. Hall BG, Acar H, Nandipati A, Barlow M. Growth rates made easy. Molecular Biology and Evolution. 2014;31: 232–238. doi:10.1093/molbev/mst187
45. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. Genome Research. 2017;27: 722–736. doi:10.1101/gr.215087.116
46. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. Phillippy AM, editor. PLOS Computational Biology. 2017;13: e1005595. doi:10.1371/journal.pcbi.1005595
47. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: Interactive visualization of de novo genome assemblies. Bioinformatics. 2015;31: 3350–3352. doi:10.1093/bioinformatics/btv383
48. Seemann T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics. 2014;30: 2068–2069. doi:10.1093/bioinformatics/btu153
49. Bertelli C, Laird MR, Williams KP, Lau BY, Hoad G, Winsor GL, et al. IslandViewer 4: Expanded prediction of genomic islands for larger-scale datasets. Nucleic Acids Research. 2017;45: W30–W35. doi:10.1093/nar/gkx343

50. Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, Lees JA, et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biology. 2020;21: 180. doi:10.1186/s13059-020-02090-4
51. Paradis E, Claude J, Strimmer K. APE: Analyses of phylogenetics and evolution in R language. Bioinformatics. 2004;20: 289–290. doi:10.1093/bioinformatics/btg412
52. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30: 2114–2120. doi:10.1093/bioinformatics/btu170
53. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012;9: 357–359. doi:10.1038/nmeth.1923
54. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079. doi:10.1093/bioinformatics/btp352
55. Yang R, Nelson AC, Henzler C, Thyagarajan B, Silverstein KAT. ScanIndel: A hybrid framework for indel detection via gapped alignment, split reads and de novo assembly. Genome Medicine. 2015;7: 127. doi:10.1186/s13073-015-0251-2
56. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Research. 2016;44: D457–D462. doi:10.1093/nar/gkv1070

Figures

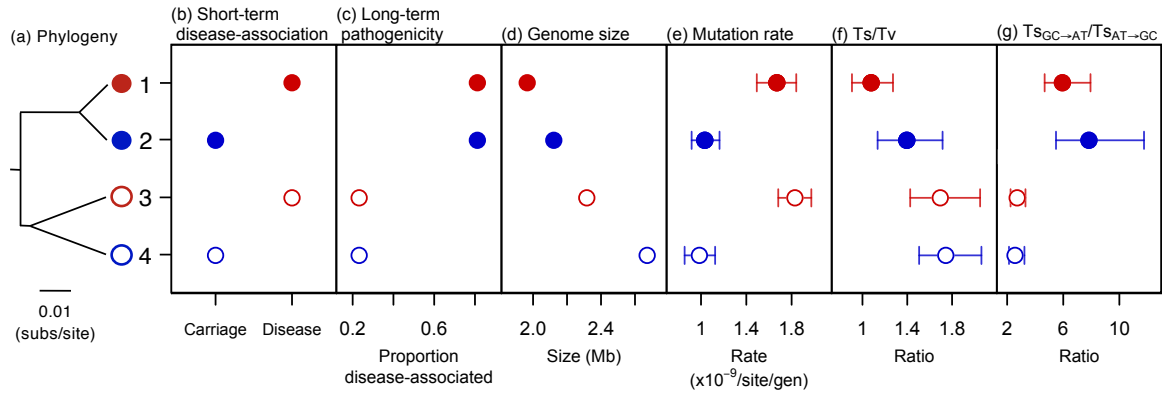


Figure 1. Differences in mutation rate between carriage and disease strains, and in mutational spectrum between more and less pathogenic groups. (a)-(d) The attributes of the strains in the 200-day MA experiment, and (e)-(g) our estimates of mutation rate and spectrum. (a) A phylogeny based on an alignment of shared genes. (b) Short-term disease association: whether the strain was isolated from the tonsils of animals without disease (carriage) or the site of an invasive infection (disease). (c) Long-term differences in pathogenicity between the two main groups of *S. suis* within which the two pairs of strains fall in a core genome phylogeny (based on an analysis of a global collection of *S. suis* isolates [14]). (d) Genome sizes. (e) Estimates of genome-wide single-base mutation rates. (f) Estimates of the ratio of transitions (Ts) to transversions (Tv). (g) Estimates of the ratio of G/C to A/T to A/T to G/C transitions. Disease strains (1 and 3) are shown in red and carriage strains (2 and 4) in blue. Strains from the more pathogenic group (1 and 2) are shown as filled circles, and strains from the less pathogenic group (3 and 4) are shown as unfilled circles. In (e)-(g) all points represent mean values across 50 replicate lines, and bars represent 95% confidence intervals estimated from bootstrapping across lines.

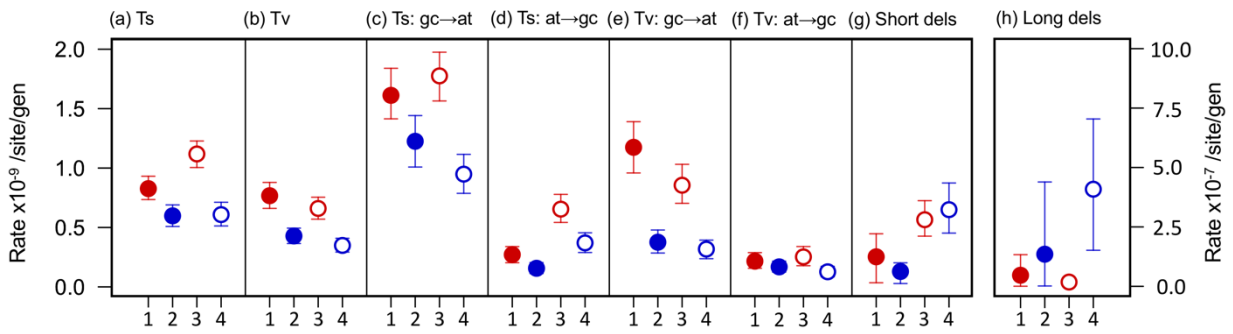


Figure 2. Disease strains have faster rates of transitions (Ts) (a), transversions (Tv) (b), G/C to A/T transitions (c), A/T to G/C transitions (d), G/C to A/T transversions (e), and A/T to G/C transversions (f), but not deletions (g, h). Disease strains are shown in red and carriage strains in blue. Strains from the more pathogenic group are shown as filled circles, and strains from the less pathogenic group are shown as empty circles. All points represent mean values across 50 replicate lines, and bars represent 95% confidence intervals estimated from bootstrapping across lines. Rates for (a)-(g) are shown on the left axis, and rates for (h) are shown on the right axis.

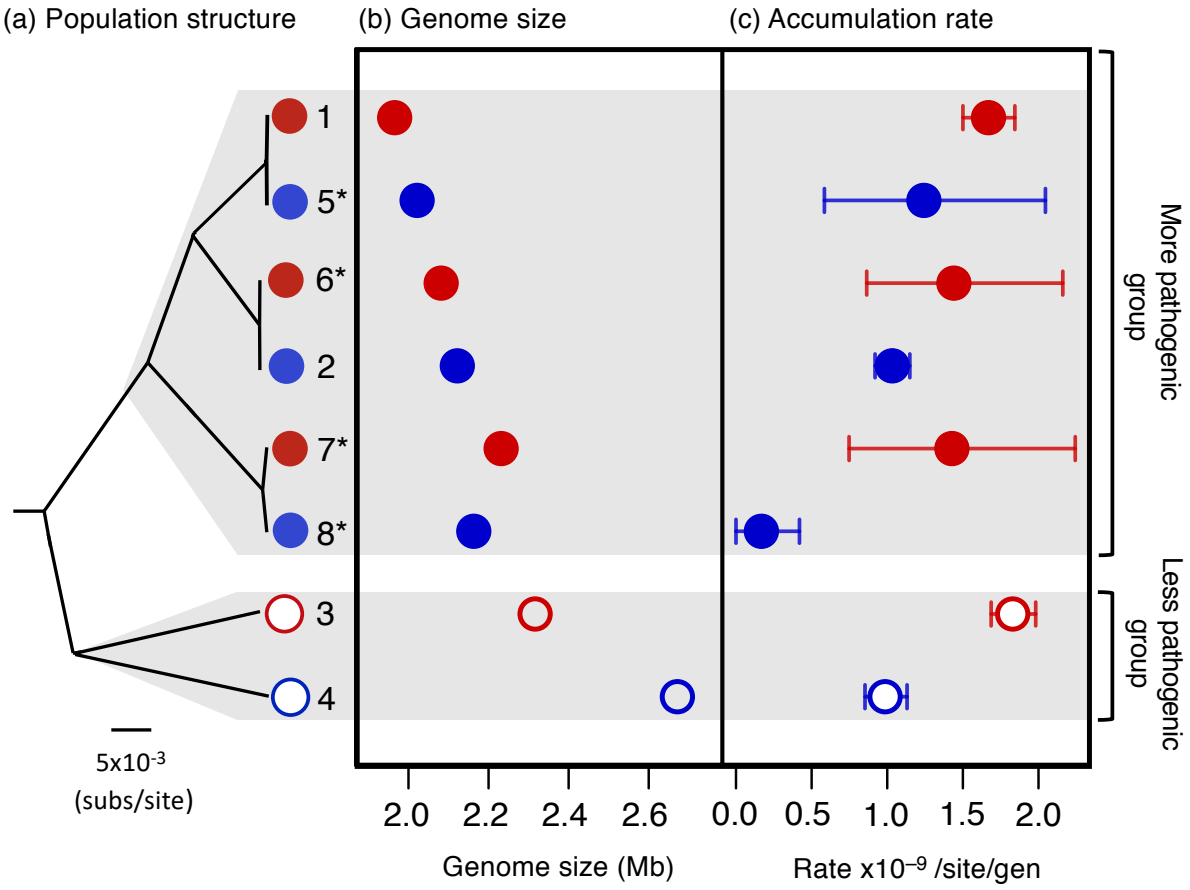


Figure 3. Disease strains have faster mutation rates compared to closely related carriage strains, independent of genome size. (a) The relationship between the eight strains in our two MA experiments based on a core gene alignment, with their location in the more and the less pathogenic groups in a phylogeny of a global collection indicated on the right (Figure S1). (b) Genome sizes of the eight strains. (c) Point estimates of the genome-wide rate of single-base mutation, based on mean values across all replicate lines. Bars represent 95% confidence intervals estimated from bootstrapping across lines. Disease strains are shown in red and carriage strains in blue. Strains from the more pathogenic group are shown as filled points, and strains from the less pathogenic group are shown as empty points. A two-sided paired t-test suggests that the difference between disease and carriage isolates is unlikely to have arisen by chance: $p = 0.035$. Strain numbers match those in Figure 1 and Table 1, and isolates from the 25-day MA experiment are indicated with *.

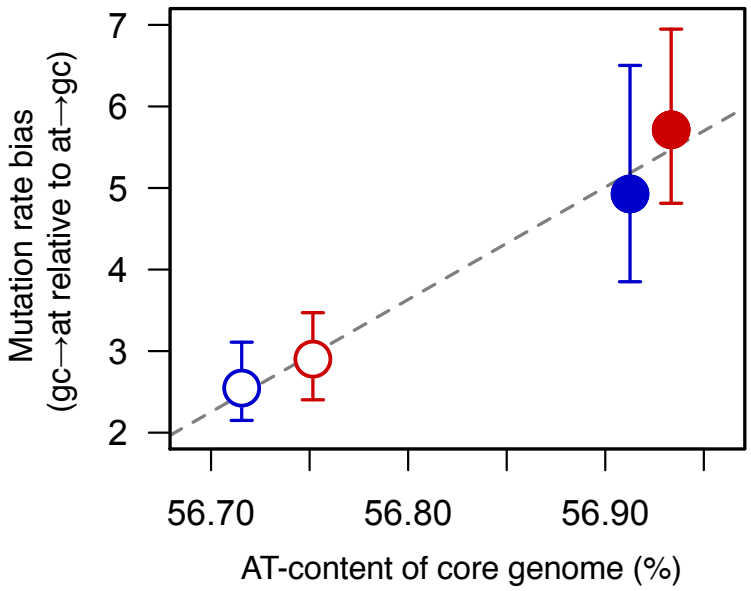
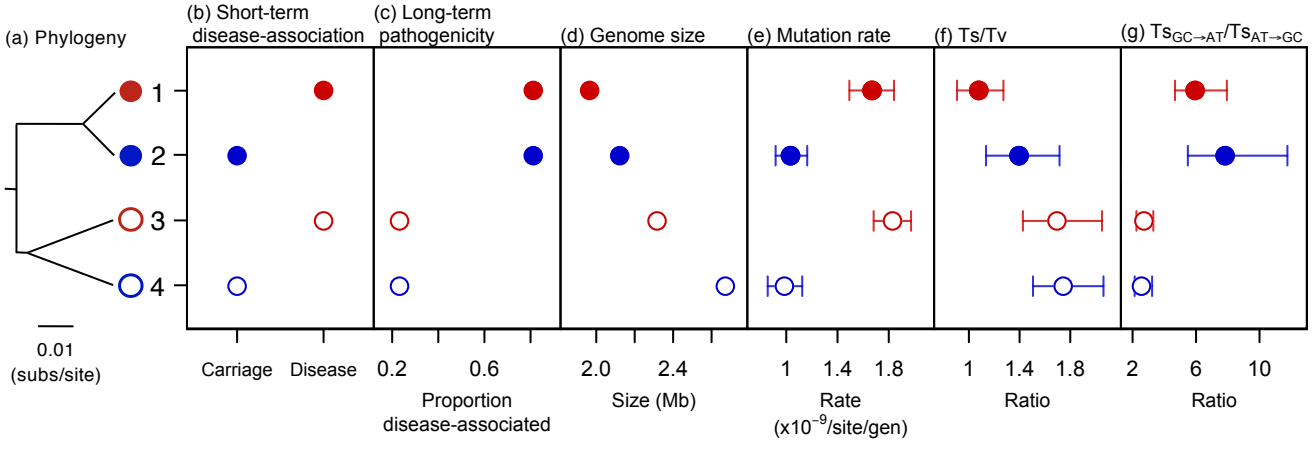
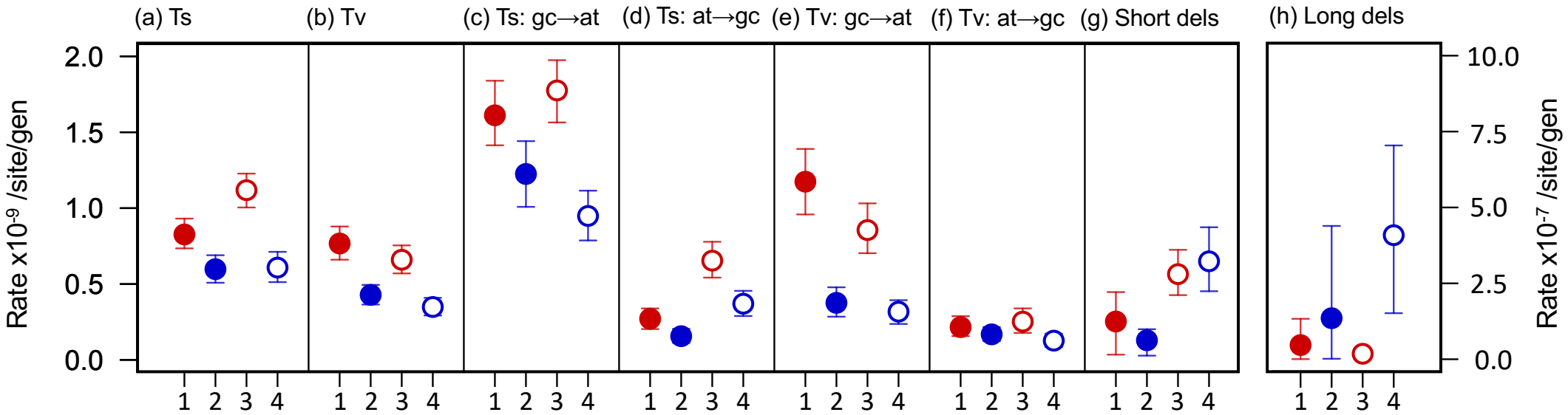


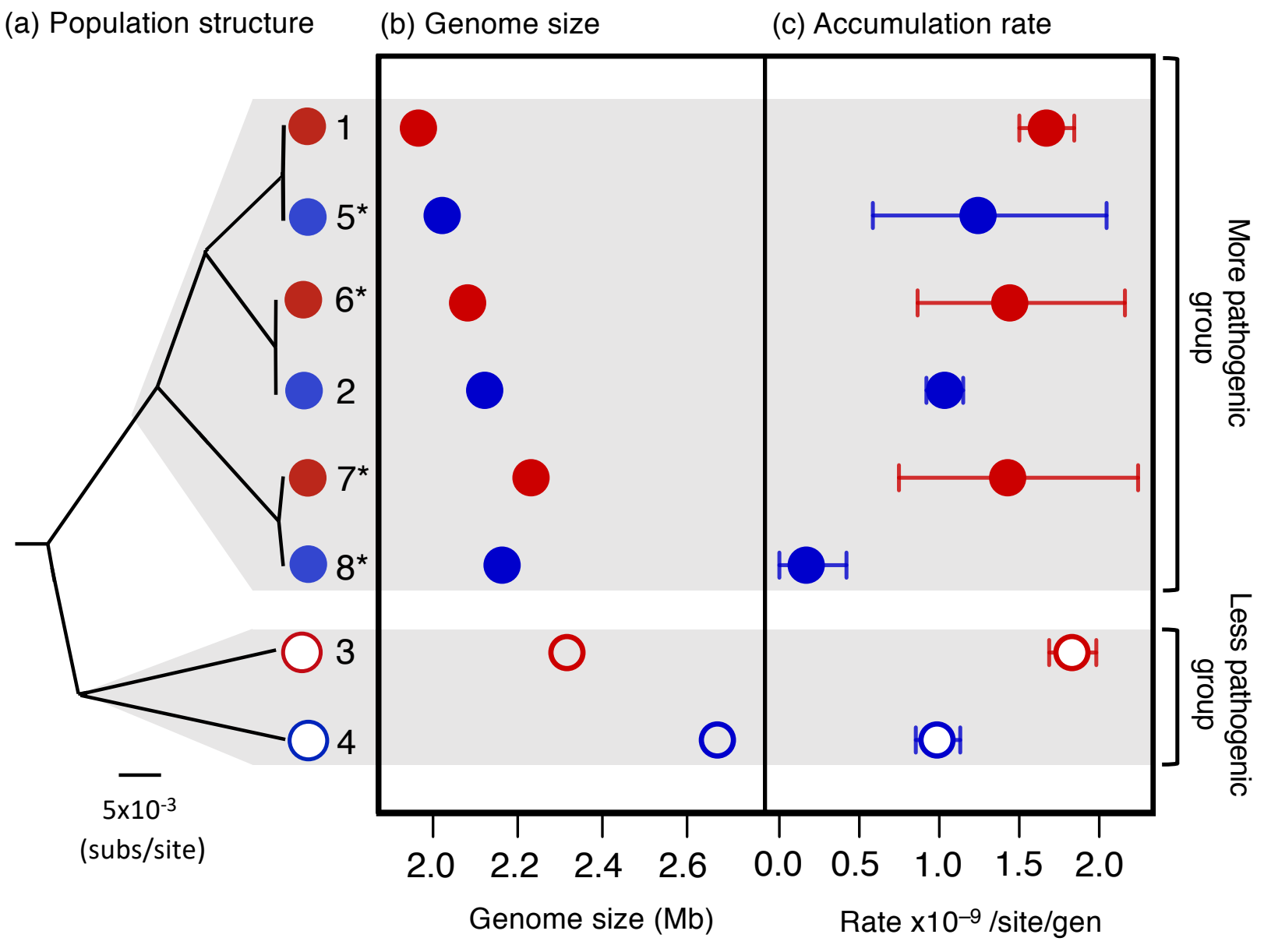
Figure 4. Mutational bias may explain differences in core genome base composition. The rate of accumulation of G/C to A/T mutations relative to the rate of accumulation of A/T to G/C mutations plotted against the proportion of bases that are A/T in genes common to all four strains from the 200-day MA experiment (1, 2, 3 and 4). All points represent mean values across 50 lines, and bars represent 95% confidence intervals estimated by bootstrapping across lines. Disease strains are shown in red and carriage strains in blue. Strains from the more pathogenic group are shown as filled circles, and strains from the less pathogenic group are shown as empty circles. The dashed line represents a line of best fit.

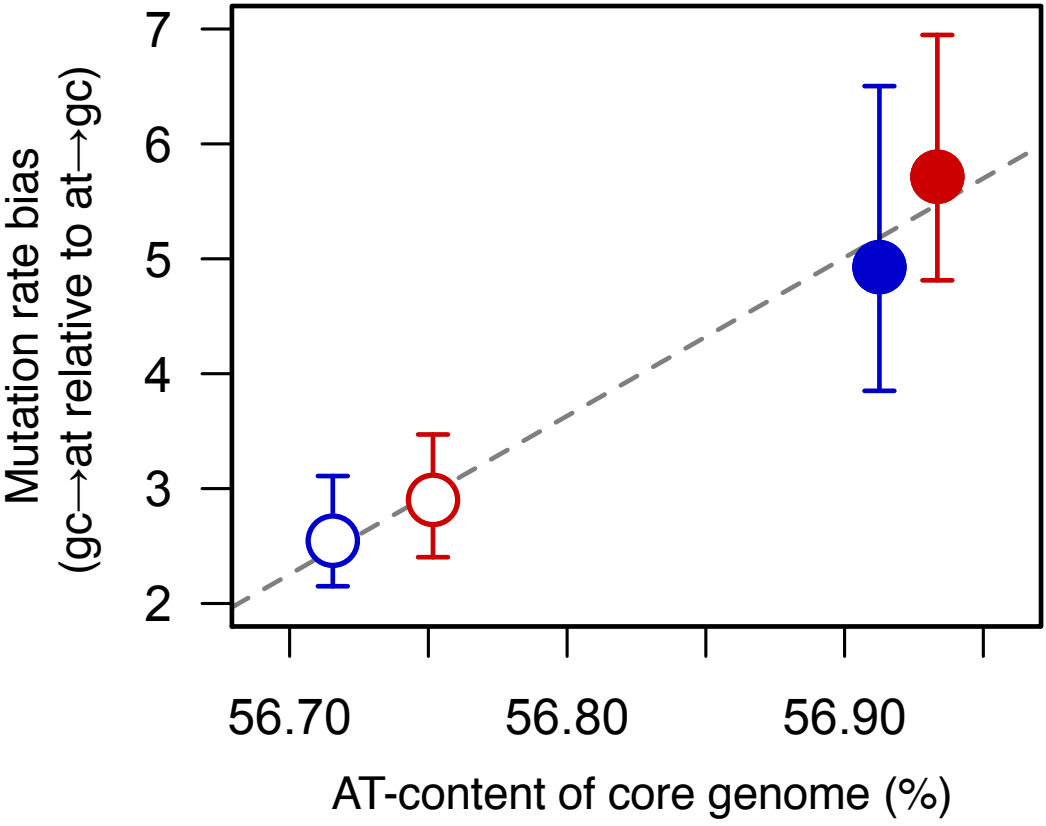
Table 1. Characteristics of the eight strains of *Streptococcus suis* used in both the 200-day (not shaded) and 25-day (shaded) MA experiments and estimates of their mutation rates.

Group	More pathogenic						Less pathogenic	
Strain	1	5	6	2	7	8	3	4
Disease-association	Disease	Carriage	Disease	Carriage	Disease	Carriage	Disease	Carriage
Genome size (Mb)	1.97	2.02	2.08	2.12	2.23	2.16	2.32	2.67
Duration of experiment (days)	200	25	25	200	25	25	200	200
Number of generations	3,991	523	513	3,989	511	503	3,484	3,445
Single-base mutation rate (95% CI) x10^{-9} /site/generation	1.67 (1.51-1.84)	1.24 (0.58-2.05)	1.44 (0.86-2.16)	1.03 (0.92-1.15)	1.43 (0.75-2.24)	0.17 (0.01-0.42)	1.83 (1.68-1.97)	0.99 (0.84 - 1.12)


(a) Phylogeny
(b) Short-term disease-association
(c) Long-term pathogenicity
(d) Genome size
(e) Mutation rate
(f) Ts/Tv
(g) $Ts_{GC\rightarrow AT}/Ts_{AT\rightarrow GC}$
1
2
3
4
0.01 (subs/site)
Carriage
Disease
0.2
0.6
Proportion disease-associated
2.0
2.4
Size (Mb)
1
1.4
1.8
Rate ($\times 10^{-9}$/site/gen)
1
1.4
1.8
Ratio
2
6
10
Ratio


(a) Ts
(b) Tv
(c) Ts: gc→at
(d) Ts: at→gc
(e) Tv: gc→at
(f) Tv: at→gc
(g) Short dels
(h) Long dels
Rate x10-9 /site/gen
Rate x10-7 /site/gen
2.0
1.5
1.0
0.5
0.0
10.0
7.5
5.0
2.5
0.0
1 2 3 4


(a) Population structure
(b) Genome size
(c) Accumulation rate
1
5*
6*
2
7*
8*
3
4
5x10-3
(subs/site)
2.0
2.2
2.4
2.6
Genome size (Mb)
0.0
0.5
1.0
1.5
2.0
Rate x10−9 /site/gen
More pathogenic group
Less pathogenic group


Mutation rate bias
(gc→at relative to at→gc)
7
6
5
4
3
2
56.70
56.80
56.90
AT-content of core genome (%)