Mutation rate dynamics reflect ecological changes 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, CB3 0ES, UK.
2. Current affiliation: School of Pharmacy and Biomedical Science, University of Portsmouth, PO1 U2P, UK.
3. Current affiliation: Cancer Biomarker Centre, Cancer Research UK Manchester Institute, The University of Manchester, Alderley Park, SK10 4TG, UK.
4. Département de Pathologie et Microbiologie, Université de Montréal, Canada.
5. Current affiliation: Haverford College, Pennsylvania, 19041, USA.

† Corresponding author: ggrmurray@gmail.com

* Joint senior authors

Abstract

While mutation is often deleterious, it can also be adaptive. Mutation rates are therefore subject to a trade-off, and this might vary with both ecology and genome size. As bacterial pathogens must survive in challenging environments and often undergo genome reduction, both factors might lead them to evolve higher mutation rates. To investigate these predictions, 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 of ecology 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 (1–4). 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 (5–9).

Bacterial pathogens face challenging and hostile environments, and also tend to have smaller genomes with fewer genes than closely related non-pathogens (10, 11). Both factors might lead pathogens to evolve higher mutation rates. Ecologies that involve growth in challenging or variable environments demand frequent adaptation, and may therefore benefit from higher mutation rates (4, 5, 7). In addition, organisms with smaller genomes experience fewer mutations per generation for the same mutation rate per site, and this means that a higher mutation rate may be associated with a lower selective cost (12, 13). Despite these predictions, the influence of ecology and genome size on mutation rate dynamics in 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, 14). Within several bacterial species, hypermutable strains have been identified that have orders of magnitude higher mutation rates than the species average due to the loss or pseudogenisation of genes in DNA repair pathways (2, 8, 15–21). While there is evidence that these rate elevations promote adaptation to environmental challenges, their association with specific ecologies remains unclear. Furthermore, hypermutation is less likely than smaller-scale changes in rate to contribute to long-term evolutionary dynamics.

To investigate the influence of ecology and genome size on mutation rate variation, we carried out mutation accumulation (MA) experiments and whole-genome sequencing on eight isolates of *Streptococcus suis*. This approach allowed us to obtain precise estimates of mutation rates and therefore to investigate small-scale within-species rate variation (22, 23). *S. suis* is an opportunistic and emerging zoonotic bacterial pathogen that colonises the upper respiratory tract of pigs, and causes severe invasive infections in both pigs and humans (24, 25). 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) (11, 26).

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 1a 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 (11). 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” clade) and carriage isolates more common in others (“less pathogenic” clades). One of our pairs of isolates was sampled from this more pathogenic clade (isolates 1 and 2) and the other from a less pathogenic clade (isolates 3 and 4). Our choice of strains therefore allows us to discriminate between three possible correlates of mutation rate: short-term transitions from carriage to disease, long-term changes in pathogenicity, and genome size.

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 reference 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 clades, or a correlation with genome size (Figure 1c, Tables 1 and S3). The increased 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 (Figure 2), and across both core and accessory genes (Figure S3).

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 S4). 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 S5, S6, 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 confirm 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 clade from our original experiment, and a more distantly related disease/carriage pair, also from the more pathogenic clade (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 (a paired t-test suggests that this pattern is unlikely to have arisen by chance: *p* = 0.03, 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 S7).

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 clades there is a difference in the mutational spectrum. The two isolates from the more pathogenic clade 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 1d,e). 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 clade as in previous studies (27, 28), we do not observe this in the two isolates from the more pathogenic clade (Figure S8).

The differences in mutational bias across the different clades are observed consistently across both core and accessory genes (Figure S9). 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 clade are further from their equilibrium base composition (Figure S10). This suggests that the change in bias occurred in an ancestor of the two isolates from the pathogenic clade, 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 clade, without a corresponding increase in the rate of insertion (Figure 2g, Figure S11, Table S9). As most small deletions occur in intergenic regions, which tend to be shorter in the two isolates from the more pathogenic clade, 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 clade, 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 S12, 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 clades. 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 the more/less pathogenic clade (Table S14).

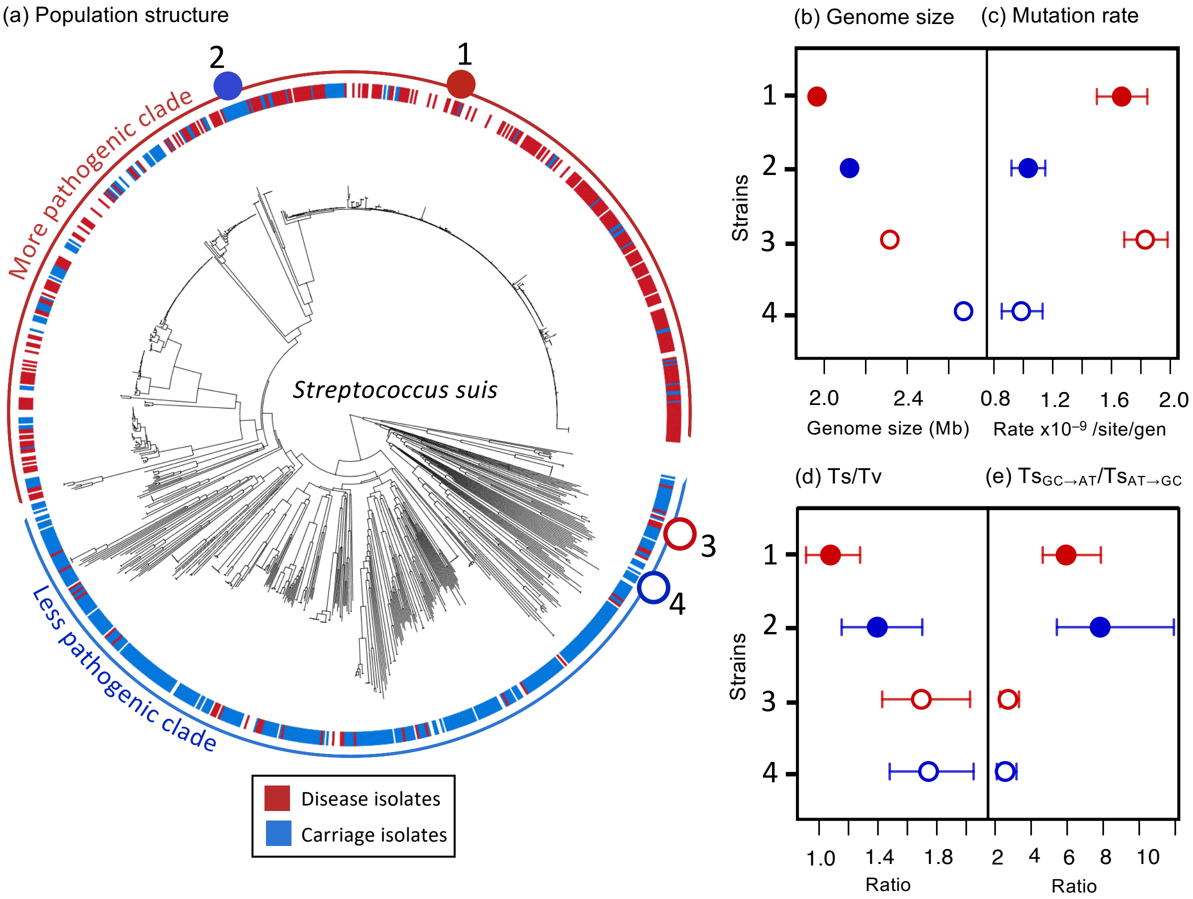


Figure 1. Differences in mutation rate between carriage and disease strains, and in mutational spectrum between more and less pathogenic clades. (a) The population structure of *S. suis*, based on an alignment of core genes from a global collection isolates (11). The coloured strip indicates whether isolates were associated with disease (red) or carriage (blue), and the outer ring indicates the location of the more pathogenic and less pathogenic clades. The four strains used in the 200-day MA experiment are indicated by points on the outer ring. (b) Genome sizes of the four strains. (c) Estimates of genome-wide single-base mutation rates. (d) Estimates of the ratio of transitions (Ts) to transversions (Tv). (e) 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 clade (1 and 2) are shown as filled circles, and strains from the less pathogenic clade (3 and 4) are shown as unfilled circles. In (c), (d) and (e) all points represent mean values across 50 replicate lines, and bars represent 95% confidence intervals estimated from bootstrapping across lines.

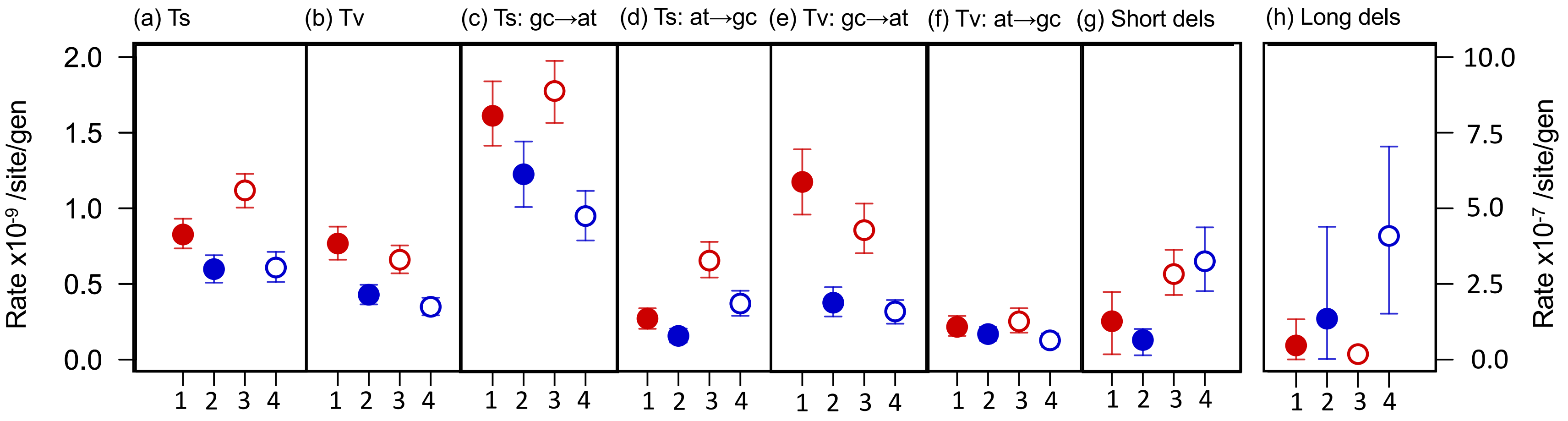


Figure 2. Disease strains have faster rates of transitions (Ts) (a), transversions (Tv) (b), G/C to A/T transitions (c), A/T to G/C transitions (d), G/C to A/T transversions (e), and A/T to G/C transversions (f), but not deletions (g, h). Disease strains are shown in red and carriage strains in blue. Strains from the more pathogenic clade are shown as filled circles, and strains from the less pathogenic clade 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.

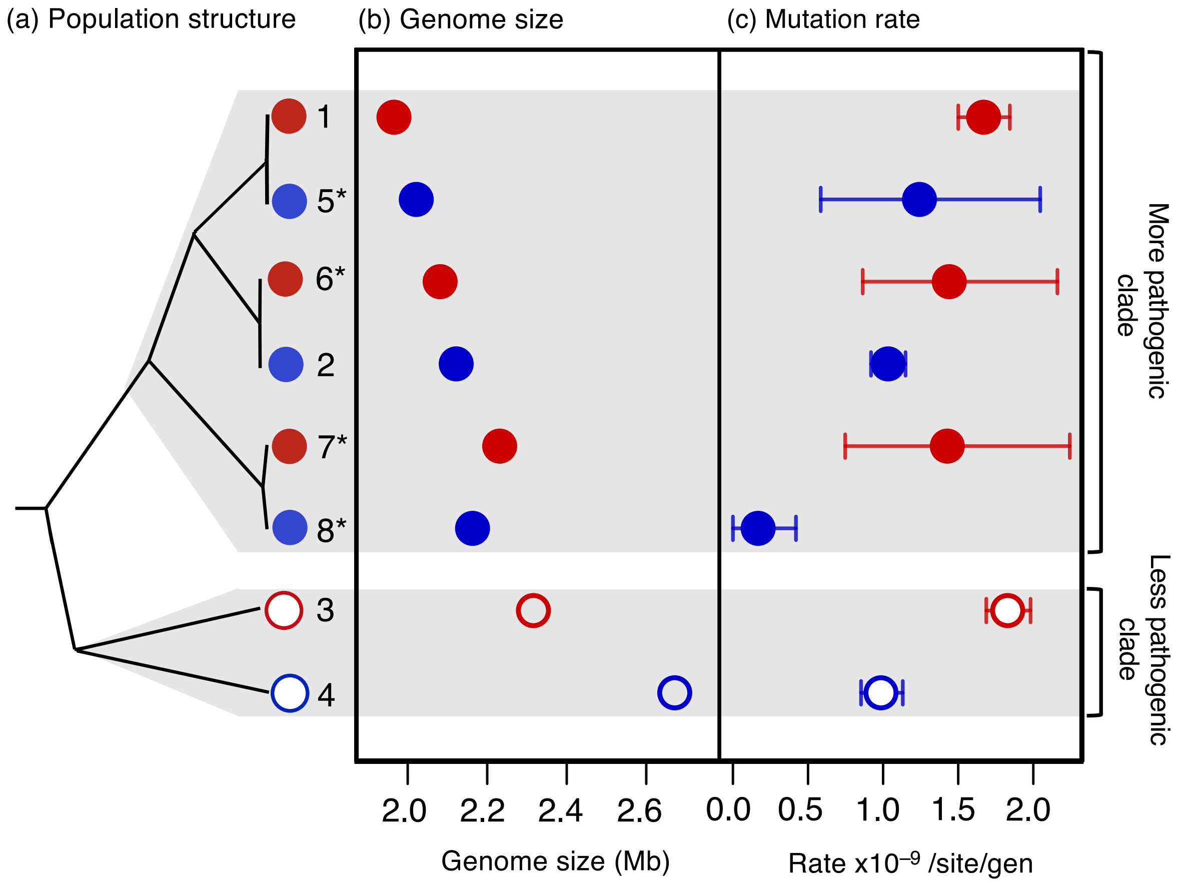


Figure 3. Disease strains have faster mutation rates compared to closely related carriage strains, independent of genome size. (a) The relationship between the eight strains in our two MA experiments based on a core gene alignment, with their location in the more and the less pathogenic clades in a phylogeny 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 clade are shown as filled points, and strains from the less pathogenic clade are shown as empty points. Strain numbers match those in Figure 1 and Table 1, and isolates from the 25-day MA experiment are indicated with *.

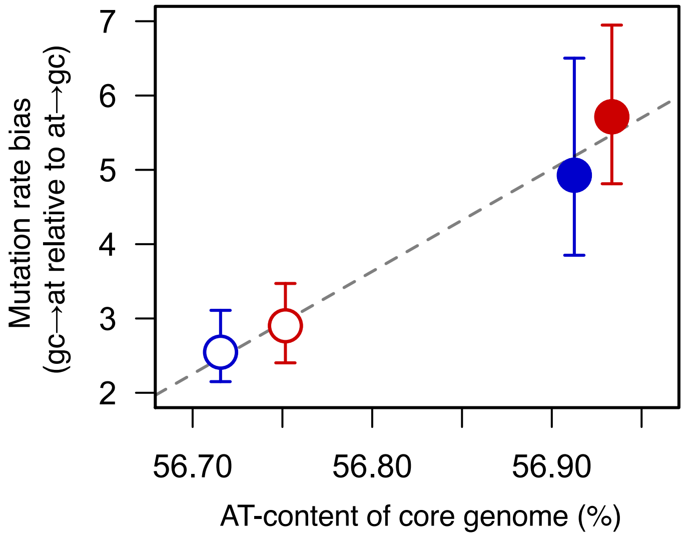


Figure 4. Mutational bias may explain differences in core genome base composition. The rate of accumulation of G/C to A/T mutations relative to the rate of accumulation of A/T to G/C mutations plotted against the proportion of bases that are A/T in genes common to all four strains from the 200-day MA 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 clade are shown as filled circles, and strains from the less pathogenic clade 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. Rate estimates are mean values across all evolved lines, with 95% confidence intervals estimated from bootstrapping across lines.

Clade	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)

Discussion

Our results revealed two associations between the pathogenic ecology of a bacterial isolate 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 clade of *S. suis* show a greater bias towards both G/C to A/T mutations and towards transversions than those from a less pathogenic clade. These changes in bias are uncoupled from disease/carriage status, and from overall mutation rate. In contrast, while genome size is a strong predictor of mutation rate variation across bacterial species (Figure S7) (1, 14), 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.

The transition from asymptomatic carriage to invasive disease is likely to involve both growth in novel environments, and additional pressures from host immune systems (29). Under these conditions, isolates with faster mutation rates may have a selective advantage because they are more likely to generate novel adaptive variants (4, 8, 30, 31), or because mutation rate is linked to a trait that is adaptive in these new conditions (e.g. 32, 33). In contrast to previous studies that have identified hypermutable strains, the rate variation we observed 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. Therefore, isolates with these higher rates may be maintained for longer, and reach higher frequencies in a population, than isolates with more extreme rate elevations. While the strains in our experiments only represent a single isolate from any given host, the difference we observe between carriage and disease isolates plausibly reflects within-host evolution, driven by selection on standing variation in carriage populations. Several studies have identified much greater variation in mutation rates in within-host bacterial populations (2, 9, 18, 21, 34), and we observe no correlation between the difference in rate and the evolutionary distance between disease and carriage isolates.

In addition to the difference in point mutation rates across disease and carriage isolates, we observed a difference in deletion rates. Almost all of the large deletion events in our 200-day MA experiments 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 (11). 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 (35).

Several clusters of *S. suis* have evolved to become more pathogenic to pigs, and are also responsible for zoonotic disease in humans (11, 26). The emergence of these more pathogenic clusters is known to be associated with both genome reduction and an increased AT-richness of the core genome (11). Here, we have found evidence that it is also associated with a change in mutational spectrum. Increased AT-richness is a common correlate of genome reduction in bacterial symbionts (36–38), and has also been observed in the evolution of the bacterial 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 (36, 37). 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 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 means 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 (11) together with three isolates sampled from pig farms in Europe (Table S2). 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 (11). 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 S3 and S4 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).

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

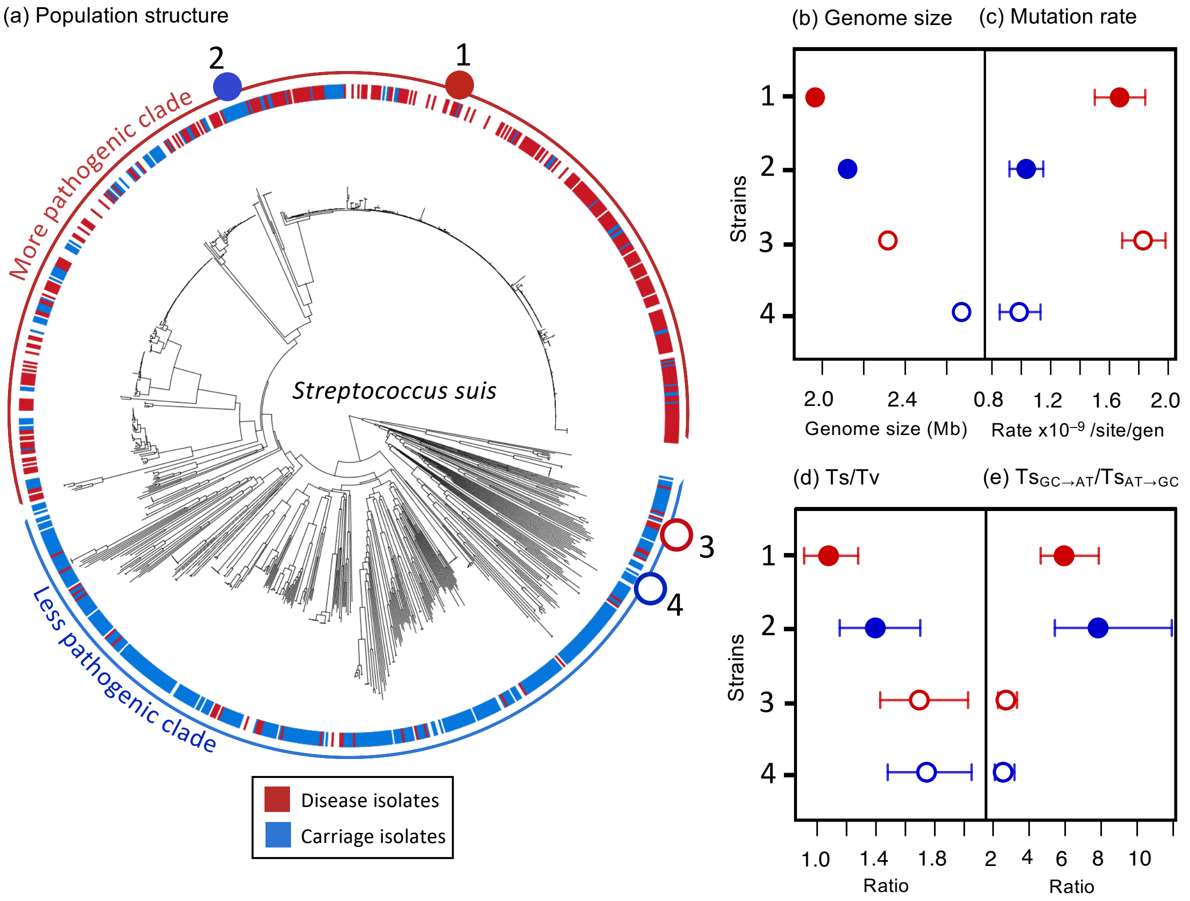
The authors declare no competing interests.

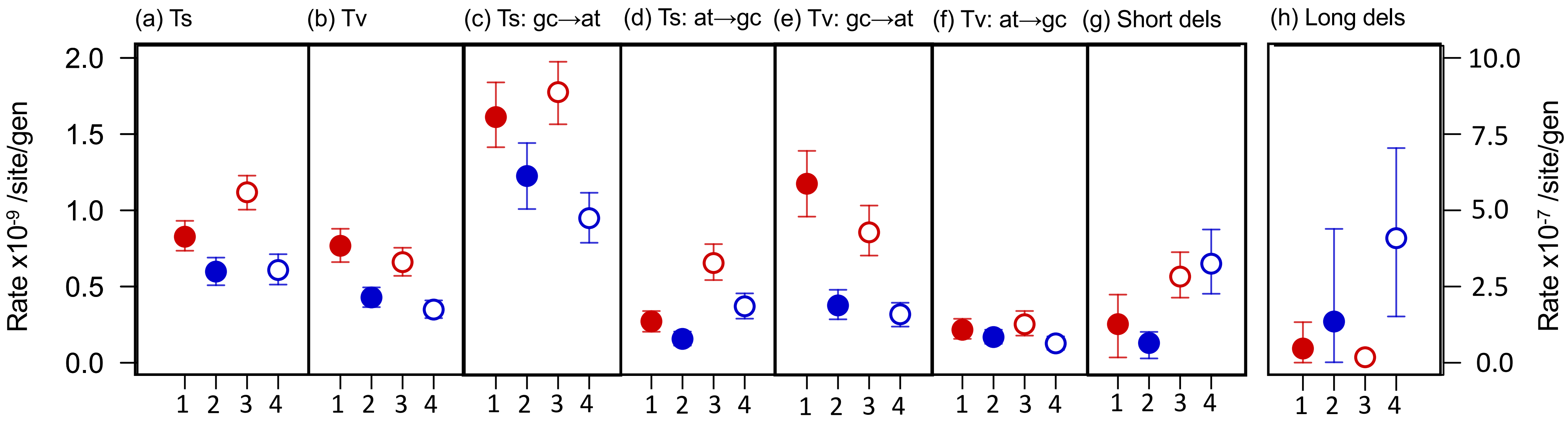
References

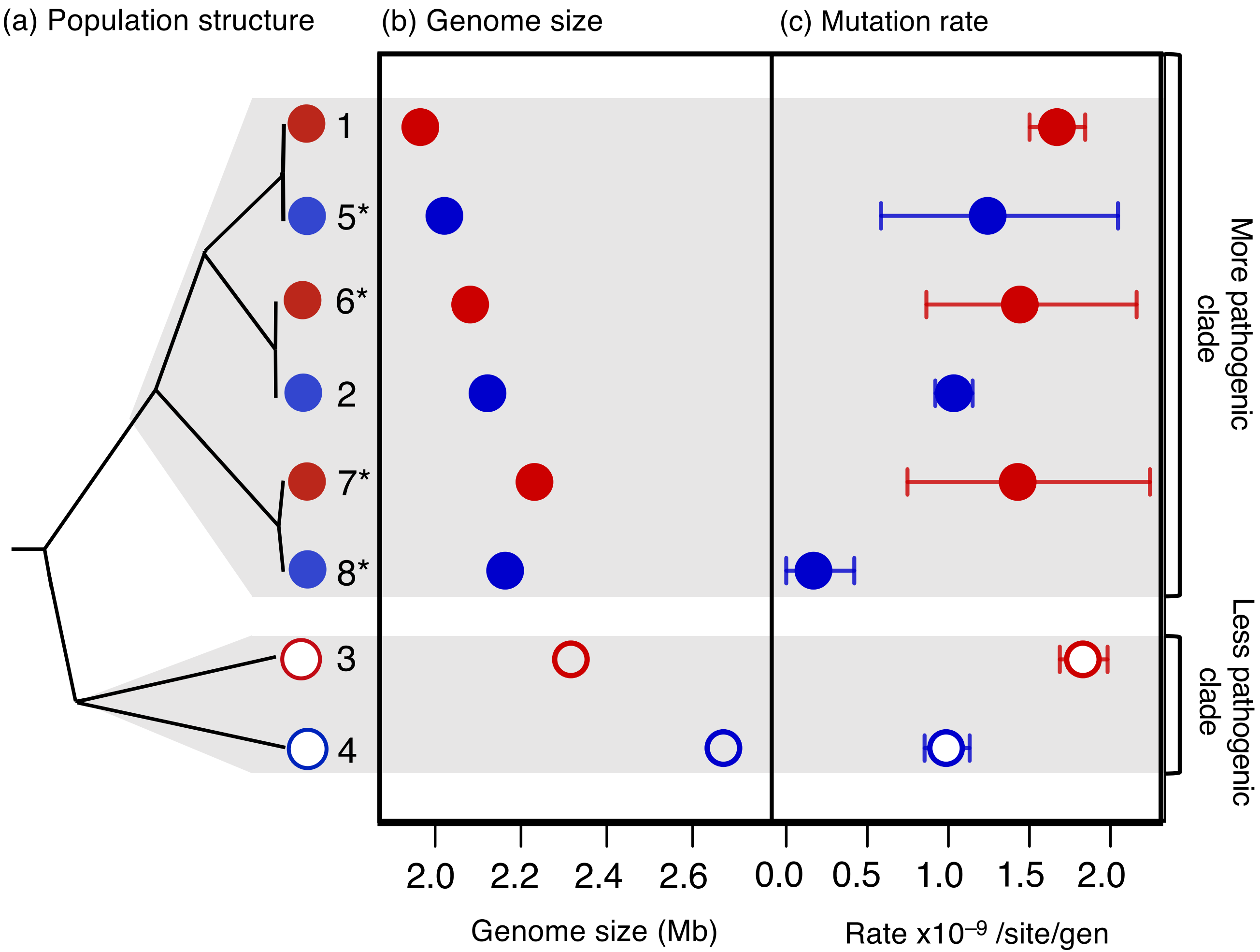
1. M. Lynch, *et al.*, Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics* **17**, 704–714 (2016).
2. R. S. Ramiro, P. Durão, C. Bank, I. Gordo, Low mutational load and high mutation rate variation in gut commensal bacteria. *PLoS Biology* **18**, e3000617 (2020).
3. E. Denamur, I. Matic, Evolution of mutation rates in bacteria. *Molecular Microbiology* **60**, 820–827 (2006).
4. A. Giraud, M. Radman, I. Matic, F. Taddei, The rise and fall of mutator bacteria. *Current Opinion in Microbiology* **4**, 582–585 (2001).
5. S. K. Sheppard, D. S. Guttman, J. R. Fitzgerald, Population genomics of bacterial host adaptation. *Nature Reviews Genetics* **19**, 549 (2018).
6. X. Didelot, A. S. Walker, T. E. Peto, D. W. Crook, D. J. Wilson, Within-host evolution of bacterial pathogens. *Nature Reviews Microbiology* **14**, 150–162 (2016).
7. C. Bonneaud, L. A. Weinert, B. Kuijper, Understanding the emergence of bacterial pathogens in novel hosts. *Philosophical Transactions of the Royal Society B: Biological Sciences* **374**, 20180328 (2019).
8. T. Swings, *et al.*, Adaptive tuning of mutation rates allows fast response to lethal stress in *Escherichia coli*. *eLife* **6** (2017).
9. A. Oliver, R. Cantón, P. Campo, F. Baquero, J. Blázquez, High frequency of hypermutable *Pseudomonas aeruginosa* in cystic fibrosis lung infection. *Science* **288**, 1251–1253 (2000).
10. L. A. Weinert, J. J. Welch, Why might bacterial pathogens have small genomes? *Trends in Ecology and Evolution* **32**, 936–947 (2017).
11. G. G. R. Murray, *et al.*, Genome reduction is associated with bacterial pathogenicity across different scales of temporal and ecological divergence. *Molecular Biology and Evolution* (2020) https:/doi.org/10.1093/molbev/msaa323.
12. S. E. Massey, The proteomic constraint and its role in molecular evolution. *Molecular Biology and Evolution* **25**, 2557–2565 (2008).
13. M. Lynch, The lower bound to the evolution of mutation rates. *Genome Biology and Evolution* **3**, 1107–1118 (2011).
14. J. W. Drake, A constant rate of spontaneous mutation in DNA-based microbes. *Proceedings of the National Academy of Sciences* **88**, 7160–7164 (1991).
15. P. R. Painter, Mutator genes and selection for the mutation rate in bacteria. *Genetics* **79**, 649–660 (1975).
16. L. Boe, *et al.*, The frequency of mutators in populations of *Escherichia coli*. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis* **448**, 47–55 (2000).
17. M. D. Gross, E. C. Siegel, Incidence of mutator strains in *Escherichia coli* and coliforms in nature. *Mutation Research Letters* **91**, 107–110 (1981).
18. J. E. LeClerc, B. Li, W. L. Payne, T. A. Cebula, High mutation frequencies among *Escherichia coli* and *Salmonella* pathogens. *Science* **274**, 1208–1211 (1996).
19. H. Wang, *et al.*, Hypermutation-induced in vivo oxidative stress resistance enhances *Vibrio cholerae* host adaptation. *PLOS Pathogens* **14**, e1007413 (2018).
20. I. Matic, *et al.*, Highly variable mutation rates in commensal and pathogenic *Escherichia coli*. *Science* **277**, 1833–1834 (1997).

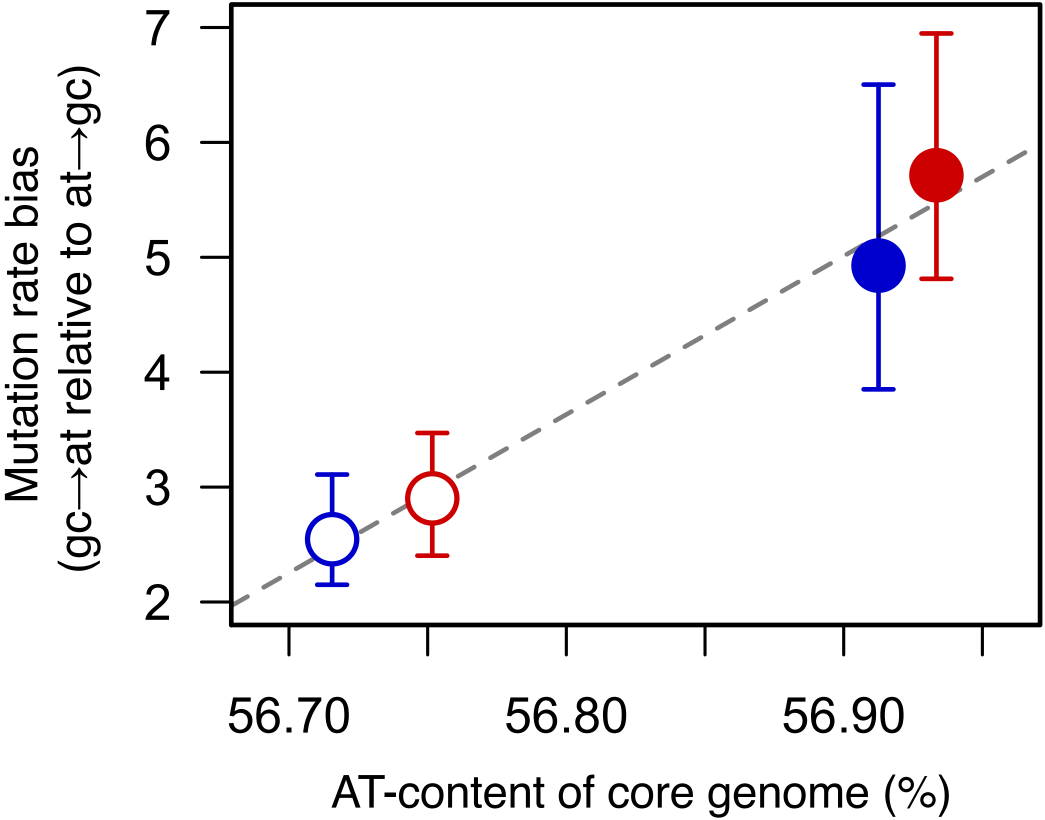
21. B. Deiham, M. Douraghi, H. Adibhesami, M. Yaseri, M. Rahbar, Screening of mutator phenotype in clinical strains of *Acinetobacter baumannii*. *Microbial Pathogenesis* **104**, 175–179 (2017).
22. D. L. Halligan, P. D. Keightley, Spontaneous mutation accumulation studies in evolutionary genetics. *Annual Review of Ecology, Evolution, and Systematics* **40**, 151–172 (2009).
23. J. E. Barrick, R. E. Lenski, Genome dynamics during experimental evolution. *Nature Reviews Genetics* **14**, 827–839 (2013).
24. D. Vötsch, M. Willenborg, Y. B. Weldearegay, P. Valentin-Weigand, *Streptococcus suis* - The “two faces” of a pathobiont in the porcine respiratory tract. *Frontiers in Microbiology* **9** (2018).
25. Z. R. Lun, Q. P. Wang, X. G. Chen, A. X. Li, X. Q. Zhu, *Streptococcus suis*: an emerging zoonotic pathogen. *Lancet Infectious Diseases* **7**, 201–209 (2007).
26. L. A. Weinert, *et al.*, Genomic signatures of human and animal disease in the zoonotic pathogen *Streptococcus suis*. *Nature Communications* **6**, 6740 (2015).
27. R. D. Blake, S. T. Hess, J. Nicholson-Tuell, The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. *Journal of Molecular Evolution* **34**, 189–200 (1992).
28. H. Lee, E. Popodi, H. Tang, P. L. Foster, 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* **109**, E2774–E2783 (2012).
29. X. Didelot, A. S. Walker, T. E. Peto, D. W. Crook, D. J. Wilson, Within-host evolution of bacterial pathogens. *Nature Reviews Microbiology* **14**, 150–162 (2016).
30. A. C. Shaver, *et al.*, Fitness evolution and the rise of mutator alleles in experimental *Escherichia coli* populations. *Genetics* **162** (2002).
31. P. D. Sniegowski, P. J. Gerrish, R. E. Lenski, Evolution of high mutation rates in experimental populations of *E. coli*. *Nature* **387**, 703–705 (1997).
32. C. Torres-Barceló, G. Cabot, A. Oliver, A. Buckling, R. C. MacLean, 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* **280**, 20130007 (2013).
33. A. R. Richardson, I. Stojiljkovic, Mismatch repair and the regulation of phase variation in *Neisseria meningitidis*. *Molecular Microbiology* **40**, 645–655 (2001).
34. A. Couce, N. Alonso-Rodriguez, C. Costas, A. Oliver, J. Blázquez, Intrapopulation variability in mutator prevalence among urinary tract infection isolates of *Escherichia coli*. *Clinical Microbiology and Infection* **22**, 566.e1-566.e7 (2016).
35. C. Howard-Varona, K. R. Hargreaves, S. T. Abedon, M. B. Sullivan, Lysogeny in nature: Mechanisms, impact and ecology of temperate phages. *ISME Journal* **11**, 1511–1520 (2017).
36. N. A. Moran, Microbial minimalism: Genome reduction in bacterial pathogens. *Cell* **108**, 583–586 (2002).
37. N. A. Moran, H. J. McLaughlin, R. Sorek, The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. *Science* **323**, 379–382 (2009).

38. H. Ochman, N. A. Moran, Genes lost and genes found: Evolution of bacterial pathogenesis and symbiosis. *Science* **292**, 1096–1099 (2001).
39. K. J. Balbi, E. P. C. Rocha, E. J. Feil, The temporal dynamics of slightly deleterious mutations in *Escherichia coli* and *Shigella* spp. *Molecular Biology and Evolution* **26**, 345–355 (2009).
40. J. B. S. Haldane, The effect of variation of fitness. *The American Naturalist* **71**, 337–349 (1937).
41. H. J. Muller, Our load of mutations. *American journal of human genetics* **2**, 111–176 (1950).
42. J. L. Payne, *et al.*, Transition bias influences the evolution of antibiotic resistance in *Mycobacterium tuberculosis*. *PLoS Biology* **17**, e3000265 (2019).
43. A. Stoltzfus, D. M. McCandlish, Mutational biases influence parallel adaptation. *Molecular Biology and Evolution* **34**, 2163–2172 (2017).
44. B. G. Hall, H. Acar, A. Nandipati, M. Barlow, Growth rates made easy. *Molecular Biology and Evolution* **31**, 232–238 (2014).
45. S. Koren, *et al.*, Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. *Genome Research* **27**, 722–736 (2017).
46. R. R. Wick, L. M. Judd, C. L. Gorrie, K. E. Holt, Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology* **13**, e1005595 (2017).
47. R. R. Wick, M. B. Schultz, J. Zobel, K. E. Holt, Bandage: Interactive visualization of de novo genome assemblies. *Bioinformatics* **31**, 3350–3352 (2015).
48. T. Seemann, Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
49. C. Bertelli, *et al.*, IslandViewer 4: Expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Research* **45**, W30–W35 (2017).
50. G. Tonkin-Hill, *et al.*, Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology* **21**, 180 (2020).
51. E. Paradis, J. Claude, K. Strimmer, APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
52. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
53. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357–359 (2012).
54. H. Li, *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
55. R. Yang, A. C. Nelson, C. Henzler, B. Thyagarajan, K. A. T. Silverstein, ScanIndel: A hybrid framework for indel detection via gapped alignment, split reads and de novo assembly. *Genome Medicine* **7**, 127 (2015).
56. M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, M. Tanabe, KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research* **44**, D457–D462 (2016).


(a) Population structure
2
1
More pathogenic clade
Streptococcus suis
3
4
Less pathogenic clade
Disease isolates
Carriage isolates
(b) Genome size
(c) Mutation rate
Strains
1
2
3
4
2.0
2.4
0.8
1.2
1.6
2.0
Genome size (Mb)
Rate x10^{-9} /site/gen
(d) Ts/Tv
(e) Ts$_{GC\rightarrow AT}$/Ts$_{AT\rightarrow GC}$
Strains
1
2
3
4
1.0
1.4
1.8
2
4
6
8
10
Ratio
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) Mutation rate
1
5*
6*
2
7*
8*
3
4
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 clade
Less pathogenic clade


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