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 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 [1–3]. 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 might benefit from higher mutation rates. Pathogens often face challenging and hostile environments, and higher mutation rates can increase the speed of adaptation [3,8,9]. 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 [10,11], and a smaller genome will mean that a higher (per site) mutation rate will lead to fewer additional mutations [12]. Third, pathogens may be less able to constrain the evolution of their mutation rate. 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 [4,13]. This could lead to a reduced efficacy of natural selection, and therefore to maladaptive evolution, including a higher mutation rate [1,14,15].

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

78 [24,25]. *S. suis* both asymptomatically colonises the upper respiratory tract of pigs, and causes
79 severe invasive infections in both pigs and humans [26,27]. Natural variation within this species
80 allows us to compare the mutation rates of closely related isolates with different pathogenic
81 ecologies and a range of genome sizes (1.97 – 2.67 Mb) [11,28].
82

83 ## Results

84 We first estimated the mutation rates of two pairs of closely related isolates, that span the known
85 range of *S. suis* genome size (strains 1-4 in Figures 1 and S1, and Tables 1 and S1). In each pair,
86 one isolate was sampled from the site of an invasive infection (disease) and the other from the
87 tonsils of a pig without *S. suis* associated disease (carriage). The relationship between these
88 isolates, and their placement in a core genome phylogeny of *S. suis* was established in a previous
89 study [11]. This study found that while both asymptomatic carriage isolates and invasive disease
90 isolates are present across the *S. suis* phylogeny, invasive disease isolates are more common in
91 one clade (a "more pathogenic" group) and carriage isolates more common outside of this clade (a
92 "less pathogenic" group). One of our pairs of isolates was sampled from this more pathogenic group
93 (isolates 1 and 2) and the other from a less pathogenic group (isolates 3 and 4) (Figure 1a and S1).
94 Our choice of strains therefore allows us to discriminate between three possible correlates of
95 mutation rate: short-term transitions from carriage to disease (Figure 1b), long-term changes in
96 pathogenicity (Figure 1c), and genome size (Figure 1d).
97

98 We estimated the mutation rates of the four isolates through four parallel 200-day MA experiments.
99 We evolved 75 replicate lines of each strain over 200 days, with daily passaging through single-
100 colony bottlenecks. We estimated that this 200-day period corresponds to between 3,445 to 3,991
101 generations for these four strains (Table 1 and S2, and Figure S2). We then sequenced the
102 genomes of 50 randomly selected evolved lines of each strain, and estimated mutation rates through
103 identifying differences in the genomes of the evolved lines compared to the genomes of the
104 ancestral strains.
105

106 **Increased mutation rates associated with the transition from carriage to disease**

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

112 transversions, and G/C to A/T transitions and transversions (Figure 2), and across both core and
113 accessory genes (Figure S3).
114
115 MA experiments aim to minimise the effect of selection on new mutations, so that accumulation rates
116 reflect mutation rates as closely as possible. Our results indicate that the influence of selection on
117 accumulation rates was low across our four MA experiments. First, distributions of single-base
118 mutations across the 50 replicate lines do not differ significantly from a Poisson distribution for any of
119 the four strains (Table S4), suggesting that accumulation rates did not vary across lines. Second,
120 rates of single-base mutation were similar across 1st and 2nd codon positions and 4-fold degenerate
121 sites for each strain (Figure S4). Third, non-unique single base mutations were rare: only 5 of the
122 2,267 single base changes observed over the four experiments occurred in more than one line, and
123 none occurred in more than two. Fourth, we found no evidence of an overall change in maximum
124 growth rates over the course of the experiment for the evolved lines of the two carriage strains or the
125 disease strain with the smaller genome, and a net decline in maximum growth rate in the disease
126 strain with the larger genome, consistent with its accumulation of the largest number of mutations,
127 and these mutations tending to have a weakly deleterious effect (Figures S5, S6, Table S5). And
128 finally, we sequenced five lines of each of the four strains at the midpoint of the experiment (100
129 days), and estimated rates of accumulation over the first and second halves of the experiment. This
130 revealed no evidence of a change in rate (Table S6).
131
132 To confirm the observed difference in mutation rate between disease and carriage isolates we
133 undertook an additional smaller scale (25-day) MA experiment with an additional four isolates. These
134 four isolates were chosen to include much closer relatives of the two isolates from the more
135 pathogenic group from our original experiment, and a more distantly related disease/carriage pair,
136 also from the more pathogenic group (Tables 1, S1 and S7). This allowed us to investigate whether
137 the difference between disease and carriage isolates holds over shorter evolutionary distances, and
138 to test the generality of the association. 15 replicate lines of the four strains were evolved over 25
139 days, 11-13 randomly selected lines of each strain were sequenced, and mutations were identified in
140 the same way as in the 200-day experiment.
141
142 Our estimates of single-base mutation rates for this combined data set of four pairs of disease and
143 carriage isolates showed a consistent pattern: disease isolates have faster rates than closely related
144 carriage isolates. While there is uncertainty associated with the rate estimates for individual strains,
145 a paired t-test suggests that the consistent difference between disease and carriage isolates we

146 observe in our point estimates is unlikely to have arisen by chance: $p = 0.03$, Figure 3, Table S8).
147 This pattern appears to be independent of genome size as it holds across pairs with variable
148 average genome sizes, and both across pairs in which the disease strain has a smaller genome than
149 the carriage isolate (3 pairs) and a pair in which the disease isolate has a larger genome than the
150 carriage isolate (Figure 3b,c). While we observe greater variation in overall mutation rates across the
151 four carriage isolates in our two experiments than across the four disease isolates, this variation is
152 not correlated with genome size (Figures 3 and S7).
153
154 **Changes in mutational spectrum associated with increased pathogenicity**
155 Over the course of our 200-day MA experiment sufficient mutations accumulated to allow us to
156 examine differences in mutational bias across isolates (Table S3). Comparing the four isolates from
157 the 200-day experiment, we found that while there is no consistent difference in the single-base
158 mutation rate across isolates from more and less pathogenic groups there is a difference in the
159 mutational spectrum. The two isolates from the more pathogenic group have a higher rate of
160 transversions relative to transitions, and a higher rate of G/C to A/T transitions relative to A/T to G/C
161 transitions (Figure 1f,g). In addition, while we observe a context-dependency of mutation rates (a
162 higher mutation rate at sites flanked by a G/C base) in the two isolates from the less pathogenic
163 group as in previous studies [29,30], we do not observe this in the two isolates from the more
164 pathogenic group (Figure S8).
165
166 The differences in mutational bias across the different groups are observed consistently across both
167 core and accessory genes (Figure S9). Moreover, the rate of G/C to A/T mutations relative A/T to
168 G/C mutations is correlated with the base composition of core genes (despite a small sample size,
169 Pearson's correlation: $r^2 = 0.98$, $p = 0.01$; Figure 4). Comparison of our estimates of mutational bias
170 and the base composition of core genes across these four strains also reveals that the two isolates
171 from the more pathogenic group are further from their equilibrium base composition (Figure S10).
172 This suggests that the change in bias occurred in an ancestor of the two isolates from the
173 pathogenic group, and may therefore be associated with their transition to a more pathogenic
174 ecology.
175
176 **Deletion rates decline with genome reduction**
177 Deletion mutations show a different pattern of variation to single-base mutations across the four
178 strains in the 200-day MA experiment. When considering short insertions and deletions (< 30 bp), we
179 observe a faster rate of deletion in the two isolates from the less pathogenic group, without a

180 corresponding increase in the rate of insertion (Figure 2g, Figure S11, Table S9). As most small
181 deletions occur in intergenic regions, which tend to be shorter in the two isolates from the more
182 pathogenic group, the difference may be due to the absence of intergenic regions that are more
183 prone to deletion in these isolates.
184
185 In contrast, larger deletions (> 100 bp) arose at a faster rate in the two carriage isolates than the two
186 disease isolates (particularly the carriage isolate from the less pathogenic group, Figure 2h). We
187 identified 34 deletions larger than 100 bp across the four strains, with 10 larger than 10 kb (Table
188 S10). Most involved the loss of regions that show signatures of being mobile genetic elements
189 (including phage-associated genes), which are more common in the isolates with larger genomes
190 (Figure S12, Table S11, S12).
191
192 **No loss of genes from DNA-repair pathways**
193 We found no evidence that the loss or truncation of genes from known DNA repair pathways led to
194 the observed differences in mutation rate or spectrum. Genes from the base excision repair,
195 nucleotide excision repair and mismatch repair pathways were uniformly present across all eight
196 strains (Table S14). In addition, no genes that were consistently absent in disease strains were also
197 consistently present in carriage strains (or vice versa) (Tables S13). Some DNA-repair pathway
198 genes were present in multiple copies in some strains, but the presence/absence of these additional
199 copies did not correlate with disease/carriage status or location in more/less pathogenic groups. In
200 addition, the majority of DNA-repair pathway genes had a constant length across the eight strains,
201 and where small length variations were observed, these did not correlate with disease/carriage
202 status or presence in a more/less pathogenic group (Table S14).
203

204 Discussion

205 While we were only able to estimate the mutation rates of a small sample of *S. suis* isolates due to
206 the time-consuming and labour-intensive nature of MA experiments, our results indicate two possible
207 links between pathogenic ecologies and the rate and spectrum of mutations. First, *S. suis* isolates
208 sampled from the site of invasive infections have consistently higher mutation rates than carriage
209 isolates sampled from the tonsils. Second, isolates from the more pathogenic group of *S. suis* show
210 a greater bias towards both G/C to A/T mutations and transversions, than those from a less
211 pathogenic group. These changes in bias are uncoupled from disease/carriage status, and from
212 overall mutation rate. In contrast, while genome size has previously been found to be a strong
213 predictor of mutation rate variation across bacterial species (Figure S7) [1,16], we found no evidence

214 of a link with mutation rates within *S. suis*. Disease isolates with a wide range of genome sizes have
215 consistent mutation rates, and while the mutation rates of carriage isolates are more variable, they
216 are not correlated with genome size.
217
218 The transition from asymptomatic carriage to invasive disease is likely to involve both growth in
219 novel environments, and additional pressures from host immune systems [31]. Under these
220 conditions, isolates with faster mutation rates may have a selective advantage because they are
221 more likely to generate novel adaptive variants [2,3,22], or because mutation rate is linked to a trait
222 that is adaptive in these new conditions [e.g. 32,33]. While the strains in our experiments only
223 represent a single isolate from any host, the difference we observe between carriage and disease
224 isolates plausibly reflects within-host evolution. If this is the case, the rapidity of the change in rates
225 suggests that it is driven by positive selection rather than random drift. This might involve selection
226 on standing variation in carriage populations; variants with higher mutation rates could be present at
227 low frequencies in carriage populations and rise to high frequencies in invasive populations. This is
228 consistent with the results of previous studies that have identified variation in mutation rates in
229 within-host bacterial populations [6,18,21,23,34], and with our observation of no correlation between
230 the difference in rate and the evolutionary distance between disease and carriage isolates.
231
232 In contrast to previous studies that have identified hypermutable strains, the rate variation we
233 observe between disease and carriage isolates is not associated with the loss or pseudogenisation
234 of genes in DNA-repair pathways, and is much smaller in scale. While this means a smaller increase
235 in the frequency of adaptive mutations, it also means a smaller increase in the frequency of
236 deleterious mutations. These higher rates may therefore be maintained for longer and reach higher
237 frequencies in a population than larger rate increases. This could contribute to the maintenance of
238 mutation rate heterogeneity within populations, which is predicted to promote faster responses to
239 selective challenges, while reducing the long-term burden of deleterious mutations [7,35].
240
241 In addition to the difference in point mutation rates across disease and carriage isolates, we
242 observed a difference in deletion rates. Almost all of the large deletion events in our 200-day MA
243 experiments occurred in the two carriage isolates, and often involved the loss of regions that contain
244 phage-associated genes. This difference in rates might reflect fewer temperate prophages in the
245 genomes of disease isolates. In *S. suis* the transition from carriage to disease is associated with a
246 reduction in genome size, and it has been suggested that this is, in part, due to the loss of mobile
247 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 [36].

Several clusters of *S. suis* have evolved to become more pathogenic to pigs, and are also responsible for zoonotic disease in humans [11,28]. 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 [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 [13,14,37], and has also been observed in the evolution of the pathogen *Shigella* [38]. 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 [13,14]. 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 [39,40].

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 [41,42]. 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.

272 ## Methods

273 **Strains and culture conditions**

274 Strains were selected from a collection of *S. suis* isolates described in [11] together with three
275 isolates sampled from pig farms in Europe (Table S2). Isolates were classified as “disease” if they
276 were recovered from systemic sites in pigs or humans with clinical signs consistent with *S. suis*
277 infection, or from the lungs of a pig with signs of pneumonia. Isolates recovered from the tonsils or
278 tracheo-bronchus of healthy pigs or pigs without signs of *S. suis* infection were classified as
279 “carriage”. Strains for both experiments were chosen based on clinical information, genome size,
280 and location in a core genome phylogeny previously described in [11]. All strains were from pigs,
281 and all “disease” isolates were associated with systemic infections.
282
283 The solid media used for all experiments was Todd-Hewitt broth (THB) supplemented with 1.5%
284 (w/v) agarose and 0.2% yeast extract (THY) (Oxoid, Basingstoke). 20% Glycerol/THB (36.4g in 1
285 litre) with 0.2% yeast extract was used for archiving and PCR, and THB with 0.2% yeast extract was
286 used for overnight growth and growth rate experiments. Cultures were streaked to single colonies on
287 solid media and incubated overnight in a static incubator at 37°C to generate stock plates. For
288 overnight cultures, an independent colony was picked from the stock plate to inoculate THB + 0.2%
289 yeast extract broth and incubated in a static incubator at 37°C.
290

291 **Mutation accumulation experiments**

292 75 replicate lines were established from each strain in the 200-day mutation accumulation (MA)
293 experiment and 15 replicate lines in the 25-day MA experiment. Each line was passaged through
294 single-colony bottlenecks by selecting the last visible independent colony in the streak (to minimise
295 selection bias). Plates were incubated at 37°C for approximately 24 hours between each transfer. All
296 passaged plates were stored at 4°C for 24 hours, to allow lines that failed to grow during passage to
297 be reset from the stored plate. In the 200-day experiment lines were archived in liquid broth every 14
298 days, and every 100 days.
299

300 **PCR**

301 For the 200-day experiment, the four strains were grown on quadrants of a single plate, and to allow
302 for errors during passaging to be corrected we developed a multiplex PCR to distinguish the strains.
303 Two sets of primer sequences for a multiplex PCR were designed using a custom Python script to
304 identify unique regions of each strain of set lengths that would resolve in electrophoresis, along with
305 a *S. suis* positive control (Table S15). To conduct the PCR, a single colony was transferred to 150 μl

306 liquid media, 5μl of which was added to 50μl sterile MiliQ water and heated at 95°C for six minutes.
307 Amplification conditions and reagent volumes are given in Table S15. We tested all lines every 14
308 days. If a mismatch was identified, lines were set back to the previous PCR run.
309
310 **Estimation of generation times**
311 Strains were streaked to single colonies from overnight cultures on THY +0.2% yeast extract agar
312 plates and incubated at 37°C for 24 hours, with a minimum of three biological replicates. Single
313 colonies were collected by excising a small disc of agar around a colony and resuspending it in 10ml
314 of PBS. The solution was serially diluted, and dilutions spread on THY +0.2% yeast extract agar plates.
315 The plates were incubated in a static incubator at 37°C and examined after 24 hours to determine the
316 number of colony forming units (CFU). Generation times were calculated by dividing log2(average
317 CFU) by the period of growth (Table S2, Figure S2).
318
319 **Growth rate measurements**
320 Growth rates were estimated for the ancestral lines and a random sample of 25 of the sequenced
321 evolved lines of each strain in the 200-day experiment, with a minimum of three biological replicates.
322 Overnight cultures were pelleted by centrifugation at 4000xg for 3 minutes to remove spent media.
323 After discarding the supernatant, the cell pellet was resuspended in fresh THY +0.2% yeast extract
324 media to a final concentration of 10^7 CFU per well. 300µl of the culture was transferred into wells and
325 incubated in a Bioscreen C (Oy Growth Curves Ab Ltd) at 37°C, with optical density (OD_{600}) measured
326 every 5 minutes for 24 hours. For 10 evolved lines (9 of strain 1 and one of strain 3) there was
327 insufficient overnight growth to reach the required starting concentration, and growth rates could not
328 be accurately measured. Maximum growth rates were calculated by taking the slope of a linear
329 regression model of $\log_2(OD_{600})$ over time, using a 30-minute sliding window to identify the period of
330 fastest growth (for details, see [43]). To test the reliability of OD as a proxy for CFU during exponential
331 growth, additional time-sampled growth curves were completed, measuring both OD and CFU for each
332 of the four ancestral strains (Table S16).
333
334 **Sequencing**
335 Illumina whole genome sequencing was undertaken for all ancestral strains, a random sample of 50
336 evolved lines of each of the strains in the 200-day MA experiments, 5 evolved lines of each of the
337 strains in the 200-day MA experiments at day 100, and 11-13 evolved lines of each of the strains in
338 the 25-day MA experiments. DNA extraction, library preparation and sequencing using a HiSeq 2500
339 instrument (Illumina, San Diego, CA, USA) was undertaken by MicrobesNG (Birmingham, UK).

340

341 **Long-read sequencing, assembly and annotation of ancestral strains**

342 We assembled high-quality reference genomes for all eight ancestral strains using methods that
343 combine short-read and long-read sequence data. For the 200-day experiments, long-read
344 sequencing library preparation was performed using Genomic-tips and a Genomic Blood and Cell
345 Culture DNA Midi kit (Qiagen, Hilden, Germany). Sequencing was performed on the Sequel
346 instrument from Pacific Biosciences using v2.1 chemistry and a multiplexed sample preparation.
347 Reads were demultiplexed using Lima in the SMRT link software
348 (https://github.com/PacificBiosciences/barcoding). Reads shorter than 2500 bases were removed
349 using prinseq-lite.pl (https://sourceforge.net/projects/prinseq/). Hybrid assemblies, using filtered
350 PacBio and Illumina reads, and preliminary assemblies of long-read data generated with Canu v1.9
351 [44], were generated with Unicycler v0.4.7 using the normal mode and default settings [45].
352 Assembly graphs were visualised and, if necessary, manually corrected with Bandage [46]. For the
353 25-day MA experiments, library preparation, short-read and long-read sequencing, and hybrid
354 assembly with Unicycler [45] were undertaken by MicrobesNG as part of their Enhanced Genome
355 Service, which uses both Illumina and Oxford Nanopore Technologies. For 4/8 strains the
356 assemblies were single-contig, and for the other four the longest contig was >98% of the total
357 assembly length (smaller contigs likely representing plasmids or mobile elements).

358

359 Genomes were annotated using Prokka v1.14.5 [47] and putative mobile elements identified using
360 IslandViewer 4 [48]. Panaroo v1.2.2 [49] was used to identify orthologous genes (using
361 recommended parameter settings) and create alignments of shared genes. Core-genome distance
362 matrices were estimated and a neighbour-joining tree created using these alignments and the *ape*
363 package in *R* [50].

364

365 **Mutation calling in evolved lines**

366 Mutations in evolved lines were called by mapping short-read sequence data to the reference
367 genomes of the ancestral strains. Illumina reads were adapter-trimmed using Trimmomatic with a
368 sliding window quality cutoff of Q15 [51]. They were mapped to the ancestral strain (excluding any
369 short contigs) using Bowtie2, and variants called using SAMtools and BCFtools [52,53]. False
370 variant calls were identified using several approaches. First, we excluded any calls with a depth of
371 less than six reads (average coverage was always >30x) or where the reference allele was present
372 in more than 5% of reads. Second, short-read data from the ancestral strains was mapped back to
373 itself, and any variants identified were excluded. Finally, any variant calls that either had a lower

374 quality score than the maximum for the line, were within 100 bases of another variant call in any line,
375 or were within 1000 bases of the start or end of the reference genome, were checked by eye for
376 evidence of mapping error. In the 200-day experiment, clusters of mutations were identified
377 (mutations within 30 bases of another in the same line). They were common only in strain 4 (<5% of
378 mutations in the other three strains), and only 10% fell in regions of the genome shared across the 4
379 strains. These mutations were not excluded from our core analyses, but Figures S3 and S4 describe
380 the impact of their exclusion.
381
382 Indels were identified using both SAMtools and ScanIndel [53,54]. To avoid false calls, we required
383 indels to be identified by both analyses and not identified when short-read data for the ancestral
384 strain was mapped back to the reference assembly. In addition, we required that the alternate allele
385 is supported by at least five reads, at least one in each orientation, and fewer than 5% of reads
386 supporting the reference allele, a quality score in SAMtools of at least 20, and an 'IMF' of at least
387 0.8. Deletions longer than 100 bases were called by identifying extended regions with zero coverage
388 in our mapped assemblies, and confirmed through examination by eye.
389
390 **Estimating average rates and statistical comparisons**
391 The single-base mutation rate per site per generation for each strain was estimated as:
392
393 $$\mu_{bs} = \frac{m}{ng}$$
394
395 where m is the number of observed single-base mutations, n is the number of nucleotide sites
396 analysed (genome length multiplied by the number of lines), and g is the total number of generations
397 over the duration of the experiment (see above). This value was estimated for the whole genome, for
398 different subsets of sites, and for different types of single base mutations, with the number of
399 nucleotide sites analysed adjusted in each case. 95% confidence intervals were generated for
400 estimates using two approaches. First, based on 1000 bootstrap samples of the lines of each strain,
401 and second based on estimates of the standard deviation in the rate across lines. We found that
402 confidence intervals were similar across the two methods, and none of our results were contingent
403 on the use of either method.

404

405 **Identification of DNA-repair genes**

406 DNA-repair genes were identified from the Prokka and Panaroo annotations using descriptions of
407 genes in the *S. suis* DNA mismatch repair, base excision repair and nucleotide excision repair
408 pathways from KEGG [55].

409

410 **Acknowledgements**

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

422

423 **Author Contributions**

424 LAW, ELM and NFH designed the 200-day MA experiment. ELM, NFH, AJB and CLK conducted the
425 200-day MA experiments, with assistance from MM, MH and LAW. LAW, GGRM and JH designed
426 the 25-day MA experiments. JH conducted the 25-day MA experiments. AJB conducted the growth
427 curves. SB assembled reference genomes for the 200-day MA experiments. GGRM designed the
428 mutation-calling pipeline for both experiments, and conducted the bioinformatic analysis for the 200-
429 day MA experiments. JH conducted the bioinformatic analysis for the 25-day MA experiments.
430 GGRM wrote the paper with contributions from all other authors.

431

432 **Declaration of interests**

433 The authors declare no competing interests.

434

435 **References**

436 1. Lynch M, Ackerman MS, Gout JF, Long H, Sung W, Thomas WK, et al. Genetic drift,
437 selection and the evolution of the mutation rate. Nature Reviews Genetics. Nature
438 Publishing Group; 2016. pp. 704–714. doi:10.1038/nrg.2016.104
439 2. Denamur E, Matic I. Evolution of mutation rates in bacteria. Molecular Microbiology. John
440 Wiley & Sons, Ltd; 2006. pp. 820–827. doi:10.1111/j.1365-2958.2006.05150.x
441 3. Giraud A, Radman M, Matic I, Taddei F. The rise and fall of mutator bacteria. Current
442 Opinion in Microbiology. 2001;4: 582–585. doi:10.1016/S1369-5274(00)00254-X
443 4. Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. Within-host evolution of bacterial
444 pathogens. Nature Reviews Microbiology. 2016;14: 150–162.
445 doi:10.1038/nrmicro.2015.13
446 5. Bonneaud C, Weinert LA, Kuijper B. Understanding the emergence of bacterial
447 pathogens in novel hosts. Philosophical Transactions of the Royal Society B: Biological
448 Sciences. 2019;374: 20180328. doi:10.1098/rstb.2018.0328
449 6. Oliver A, Cantón R, Campo P, Baquero F, Blázquez J. High frequency of hypermutable
450 Pseudomonas aeruginosa in cystic fibrosis lung infection. Science. 2000;288: 1251–
451 1253. doi:10.1126/science.288.5469.1251
452 7. Matic I. Mutation Rate Heterogeneity Increases Odds of Survival in Unpredictable
453 Environments. Molecular Cell. Cell Press; 2019. pp. 421–425.
454 doi:10.1016/j.molcel.2019.06.029
455 8. Wiser MJ, Ribeck N, Lenski RE. Long-term dynamics of adaptation in asexual
456 populations. Science. 2013;342: 1364–1367. doi:10.1126/science.1243357
457 9. Taddei F, Radman M, Maynard-Smith J, Toupance B, Gouyon PH, Godelle B. Role of
458 mutator alleles in adaptive evolution. Nature. 1997;387: 700–702. doi:10.1038/42696
459 10. Weinert LA, Welch JJ. Why might bacterial pathogens have small genomes? Trends in
460 Ecology and Evolution. Elsevier Ltd; 2017. pp. 936–947. doi:10.1016/j.tree.2017.09.006
461 11. Murray GGR, Charlesworth J, Miller EL, Casey MJ, Lloyd CT, Gottschalk M, et al.
462 Genome reduction is associated with bacterial pathogenicity across different scales of
463 temporal and ecological divergence. Molecular Biology and Evolution. 2021;38: 1570–
464 1579. doi:10.1093/molbev/msaa323
465 12. Smith DR. The mutational hazard hypothesis of organelle genome evolution: 10 years on.
466 Molecular Ecology. 2016;25: 3769–3775. doi:10.1111/mec.13742
467 13. Moran NA. Microbial minimalism: Genome reduction in bacterial pathogens. Cell. Cell
468 Press; 2002. pp. 583–586. doi:10.1016/S0092-8674(02)00665-7

14. 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
15. 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
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

502 26. Vötsch D, Willenborg M, Weldearegay YB, Valentin-Weigand P. *Streptococcus suis* - The
503 "two faces" of a pathobiont in the porcine respiratory tract. Frontiers in Microbiology.
504 Frontiers Media S.A.; 2018. doi:10.3389/fmicb.2018.00480
505 27. Lun ZR, Wang QP, Chen XG, Li AX, Zhu XQ. *Streptococcus suis*: an emerging zoonotic
506 pathogen. Lancet Infectious Diseases. Elsevier; 2007. pp. 201–209. doi:10.1016/S1473-
507 3099(07)70001-4
508 28. Weinert LA, Chaudhuri RR, Wang J, Peters SE, Corander J, Jombart T, et al. Genomic
509 signatures of human and animal disease in the zoonotic pathogen *Streptococcus suis*.
510 Nature Communications. 2015;6: 6740. doi:10.1038/ncomms7740
511 29. Blake RD, Hess ST, Nicholson-Tuell J. The influence of nearest neighbors on the rate
512 and pattern of spontaneous point mutations. Journal of Molecular Evolution. 1992;34:
513 189–200. doi:10.1007/BF00162968
514 30. Lee H, Popodi E, Tang H, Foster PL. Rate and molecular spectrum of spontaneous
515 mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing.
516 Proceedings of the National Academy of Sciences of the United States of America.
517 2012;109: E2774–E2783. doi:10.1073/pnas.1210309109
518 31. Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. Within-host evolution of bacterial
519 pathogens. Nature Reviews Microbiology. Nature Publishing Group; 2016. pp. 150–162.
520 doi:10.1038/nrmicro.2015.13
521 32. Torres-Barceló C, Cabot G, Oliver A, Buckling A, MacLean RC. A trade-off between
522 oxidative stress resistance and DNA repair plays a role in the evolution of elevated
523 mutation rates in bacteria. Proceedings of the Royal Society B: Biological Sciences.
524 2013;280: 20130007. doi:10.1098/rspb.2013.0007
525 33. Richardson AR, Stojiljkovic I. Mismatch repair and the regulation of phase variation in
526 Neisseria meningitidis. Molecular Microbiology. 2001;40: 645–655. doi:10.1046/j.1365-
527 2958.2001.02408.x
528 34. Couce A, Alonso-Rodriguez N, Costas C, Oliver A, Blázquez J. Intrapopulation variability
529 in mutator prevalence among urinary tract infection isolates of *Escherichia coli*. Clinical
530 Microbiology and Infection. 2016;22: 566.e1-566.e7. doi:10.1016/j.cmi.2016.03.008
531 35. Alexander HK, Mayer SI, Bonhoeffer S. Population Heterogeneity in Mutation Rate
532 Increases the Frequency of Higher-Order Mutants and Reduces Long-Term Mutational
533 Load. Molecular biology and evolution. 2017;34: 419–436. doi:10.1093/molbev/msw244

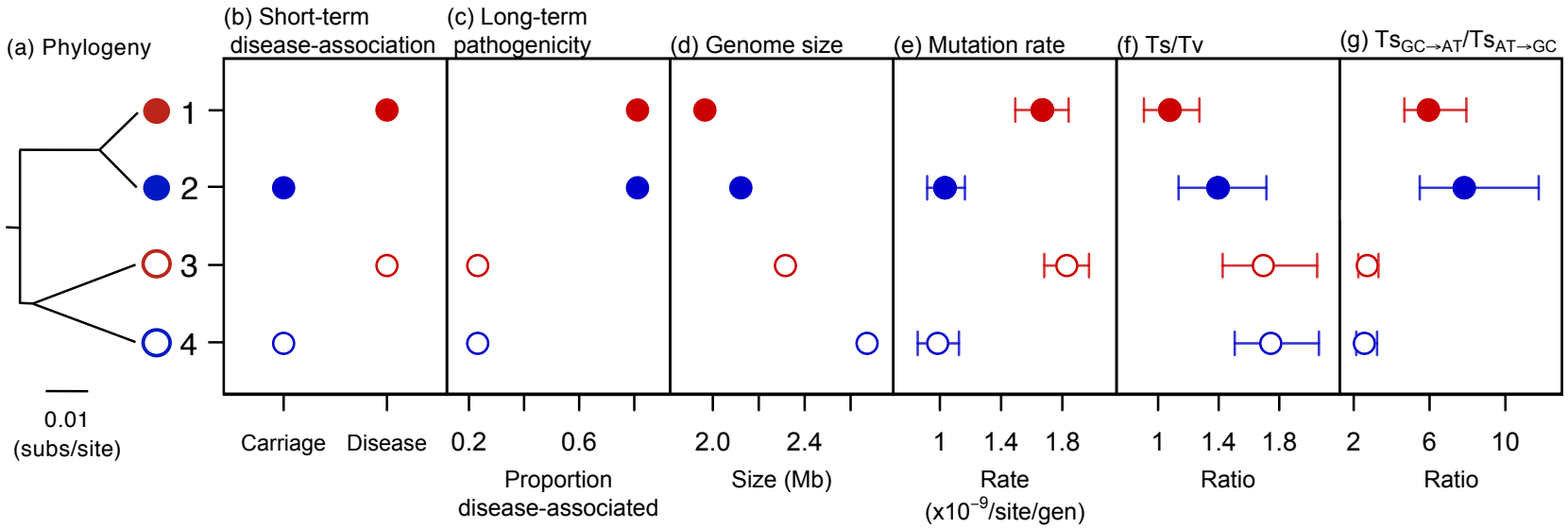
534 36. Howard-Varona C, Hargreaves KR, Abedon ST, Sullivan MB. Lysogeny in nature:
535 Mechanisms, impact and ecology of temperate phages. ISME Journal. Nature Publishing
536 Group; 2017. pp. 1511–1520. doi:10.1038/ismej.2017.16
537 37. Ochman H, Moran NA. Genes lost and genes found: Evolution of bacterial pathogenesis
538 and symbiosis. Science. 2001;292: 1096–1099. doi:10.1126/science.1058543
539 38. Balbi KJ, Rocha EPC, Feil EJ. The temporal dynamics of slightly deleterious mutations in
540 *Escherichia coli* and *Shigella* spp. Molecular Biology and Evolution. 2009;26: 345–355.
541 doi:10.1093/molbev/msn252
542 39. Haldane JBS. The effect of variation of fitness. The American Naturalist. 1937;71: 337–
543 349. doi:10.1086/280722
544 40. Muller HJ. Our load of mutations. American journal of human genetics. 1950;2: 111–176.
545 Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1716299/
546 41. Payne JL, Menardo F, Trauner A, Borrell S, Gygli SM, Loiseau C, et al. Transition bias
547 influences the evolution of antibiotic resistance in *Mycobacterium tuberculosis*. PLoS
548 Biology. 2019;17: e3000265. doi:10.1371/journal.pbio.3000265
549 42. Stoltzfus A, McCandlish DM. Mutational biases influence parallel adaptation. Molecular
550 Biology and Evolution. 2017;34: 2163–2172. doi:10.1093/molbev/msx180
551 43. Hall BG, Acar H, Nandipati A, Barlow M. Growth rates made easy. Molecular Biology and
552 Evolution. 2014;31: 232–238. doi:10.1093/molbev/mst187
553 44. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: Scalable and
554 accurate long-read assembly via adaptive κ-mer weighting and repeat separation.
555 Genome Research. 2017;27: 722–736. doi:10.1101/gr.215087.116
556 45. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome
557 assemblies from short and long sequencing reads. Phillippy AM, editor. PLOS
558 Computational Biology. 2017;13: e1005595. doi:10.1371/journal.pcbi.1005595
559 46. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: Interactive visualization of de novo
560 genome assemblies. Bioinformatics. 2015;31: 3350–3352.
561 doi:10.1093/bioinformatics/btv383
562 47. Seemann T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics. 2014;30:
563 2068–2069. doi:10.1093/bioinformatics/btu153
564 48. Bertelli C, Laird MR, Williams KP, Lau BY, Hoad G, Winsor GL, et al. IslandViewer 4:
565 Expanded prediction of genomic islands for larger-scale datasets. Nucleic Acids
566 Research. 2017;45: W30–W35. doi:10.1093/nar/gkx343

567 49. Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, Lees JA, et al. Producing
568 polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biology. 2020;21:
569 180. doi:10.1186/s13059-020-02090-4
570 50. Paradis E, Claude J, Strimmer K. APE: Analyses of phylogenetics and evolution in R
571 language. Bioinformatics. 2004;20: 289–290. doi:10.1093/bioinformatics/btg412
572 51. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence
573 data. Bioinformatics. 2014;30: 2114–2120. doi:10.1093/bioinformatics/btu170
574 52. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods.
575 2012;9: 357–359. doi:10.1038/nmeth.1923
576 53. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
577 Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079.
578 doi:10.1093/bioinformatics/btp352
579 54. Yang R, Nelson AC, Henzler C, Thyagarajan B, Silverstein KAT. ScanIndel: A hybrid
580 framework for indel detection via gapped alignment, split reads and de novo assembly.
581 Genome Medicine. 2015;7: 127. doi:10.1186/s13073-015-0251-2
582 55. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference
583 resource for gene and protein annotation. Nucleic Acids Research. 2016;44: D457–D462.
584 doi:10.1093/nar/gkv1070
585

586

587 **Figures**

588



589

590 **Figure 1. Differences in mutation rate between carriage and disease strains, and in mutational spectrum between more and less**
591 **pathogenic groups.** (a)-(d) The attributes of the strains in the 200-day MA experiment, and (e)-(g) our estimates of mutation rate and
592 spectrum. (a) A phylogeny based on an alignment of shared genes. (b) Short-term disease association: whether the strain was isolated
593 from the tonsils of animals without disease (carriage) or the site of an invasive infection (disease). (c) Long-term differences in
594 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
595 analysis of a global collection of *S. suis* isolates [14]). (d) Genome sizes. (e) Estimates of genome-wide single-base mutation rates. (f)
596 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
597 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
598 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
599 across 50 replicate lines, and bars represent 95% confidence intervals estimated from bootstrapping across lines.

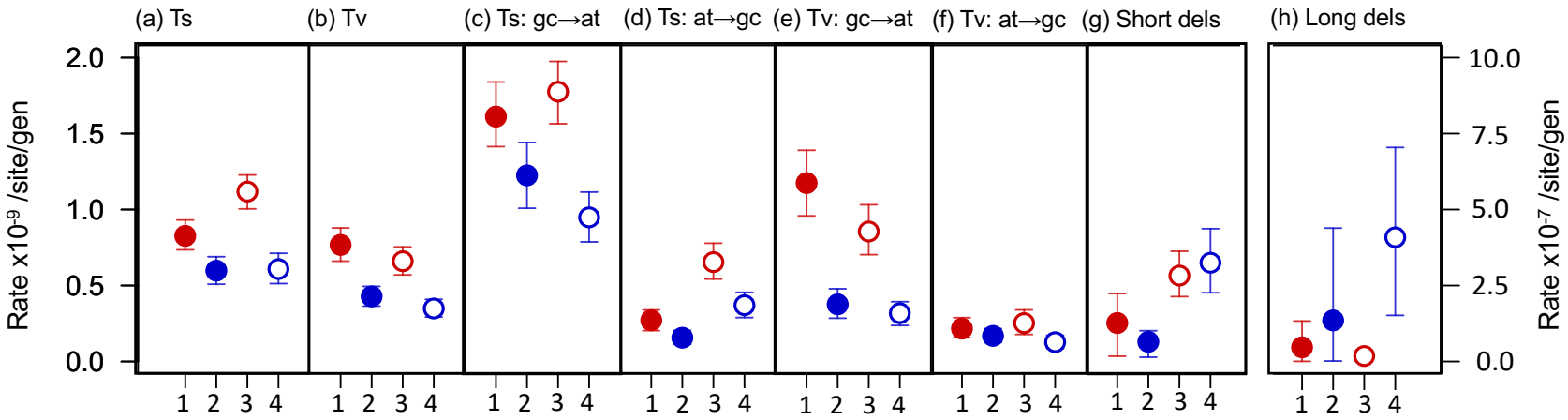


Figure 2. Disease strains have faster rates of transitions (Ts) (a), transversions (Tv) (b), G/C to A/T transitions (c), A/T to G/C transitions (d), G/C to A/T transversions (e), and A/T to G/C transversions (f), but not deletions (g, h). Disease strains are shown in red and carriage strains in blue.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

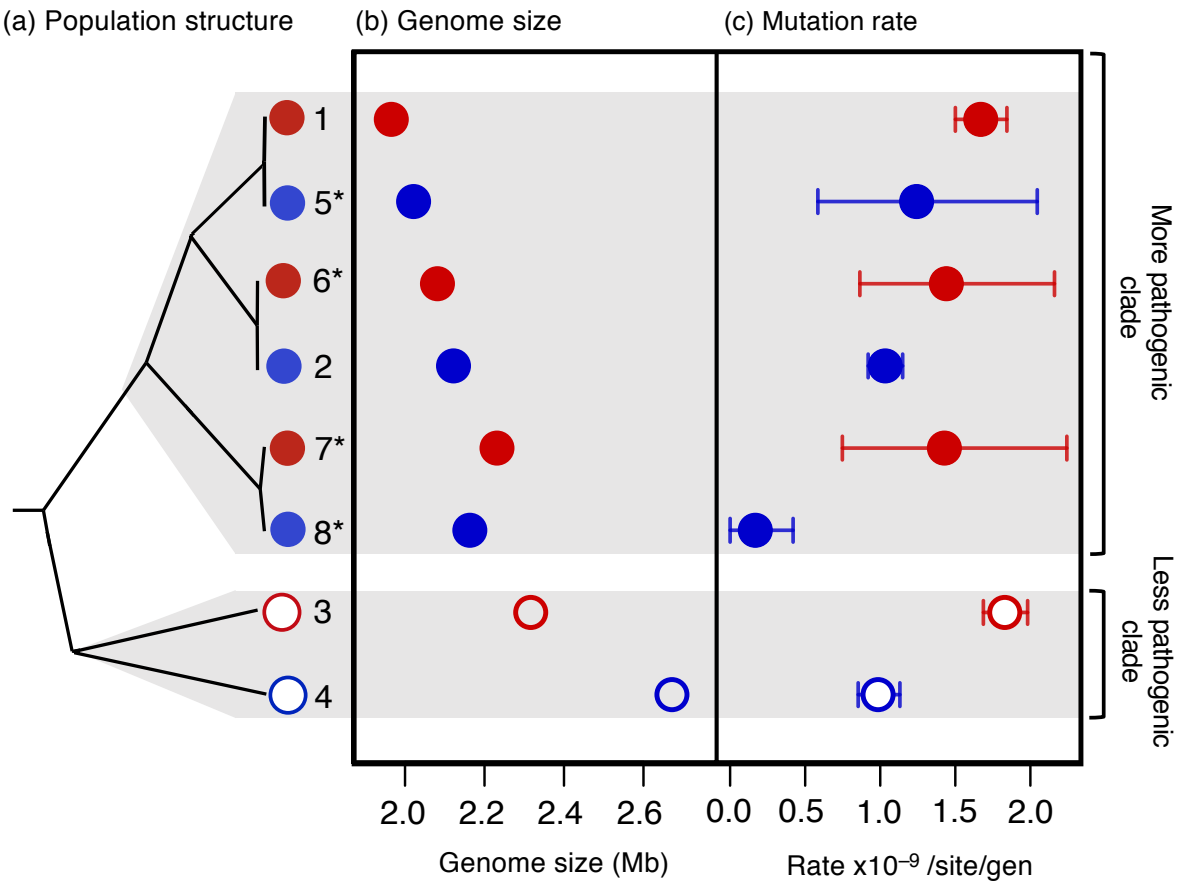
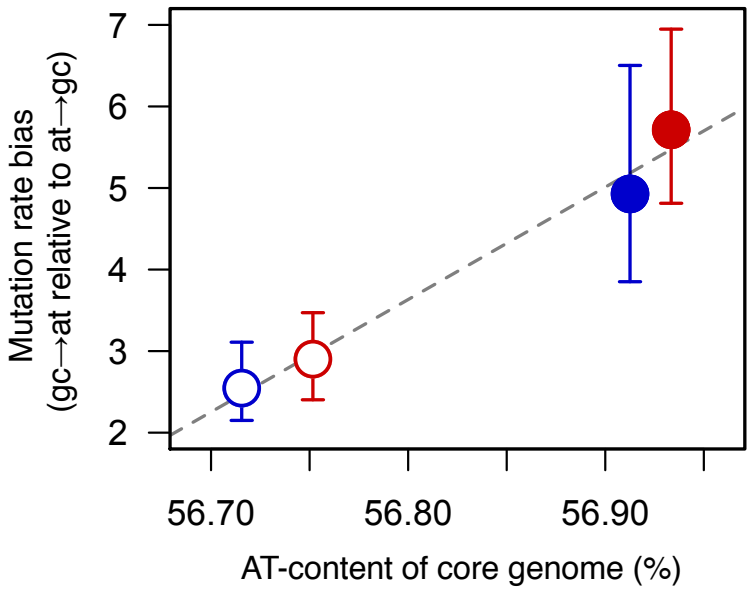


Figure 3. Disease strains have faster mutation rates compared to closely related carriage strains, independent of genome size. (a) The relationship between the eight strains in our two MA experiments based on a core gene alignment, with their location in the more and the less pathogenic groups in a phylogeny 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 paired t-test suggests that the difference between disease and carriage isolates is unlikely to have arisen by chance: $p = 0.03$. Strain numbers match those in Figure 1 and Table 1, and isolates from the 25-day MA experiment are indicated with *.

625



626

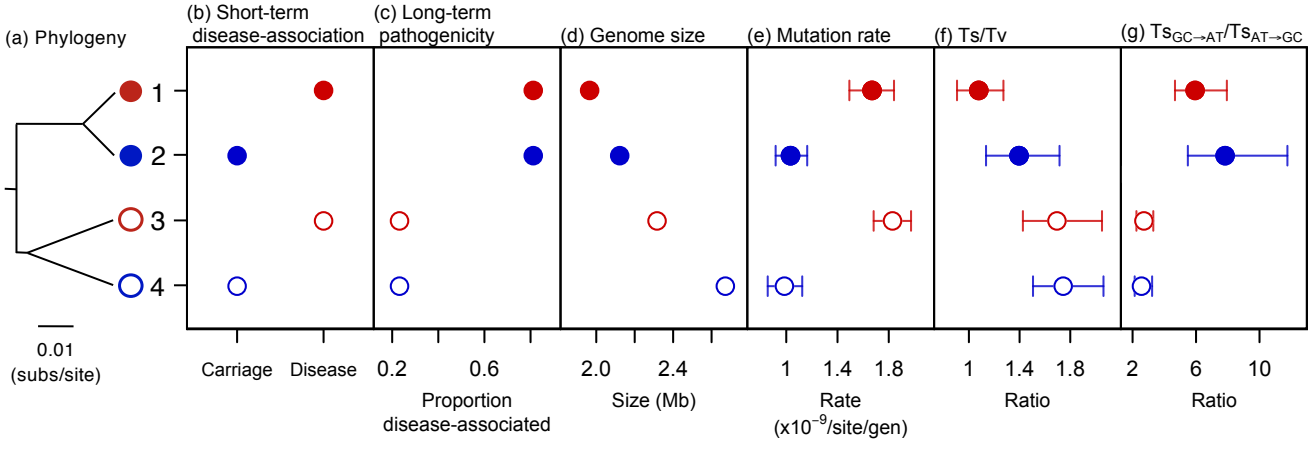
627

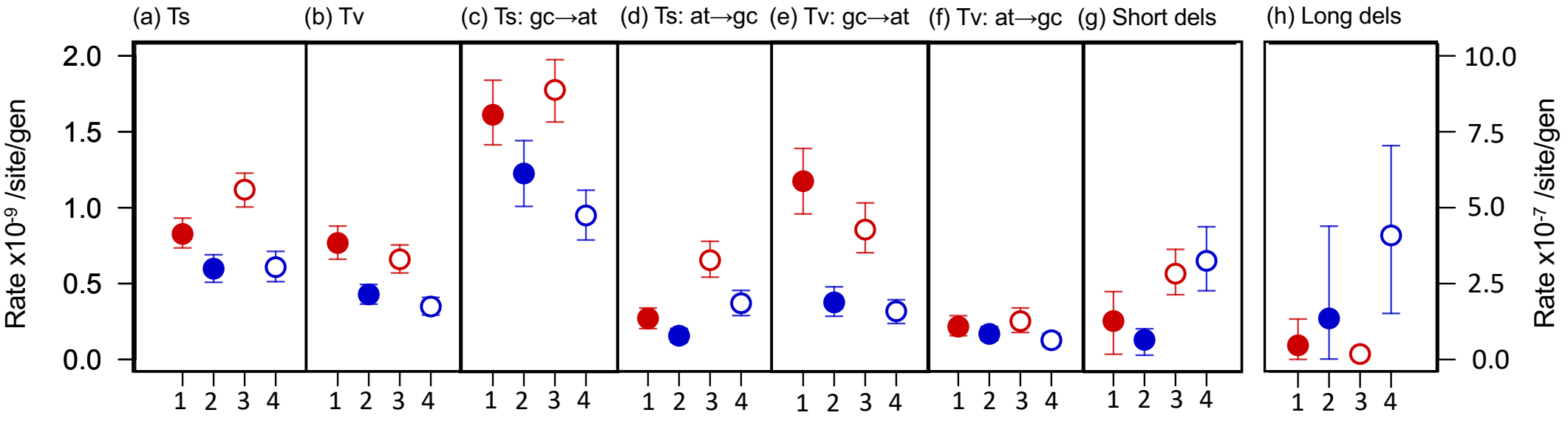
628 **Figure 4. Mutational bias may explain differences in core genome base composition.** The rate
629 of accumulation of G/C to A/T mutations relative to the rate of accumulation of A/T to G/C mutations
630 plotted against the proportion of bases that are A/T in genes common to all four strains from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

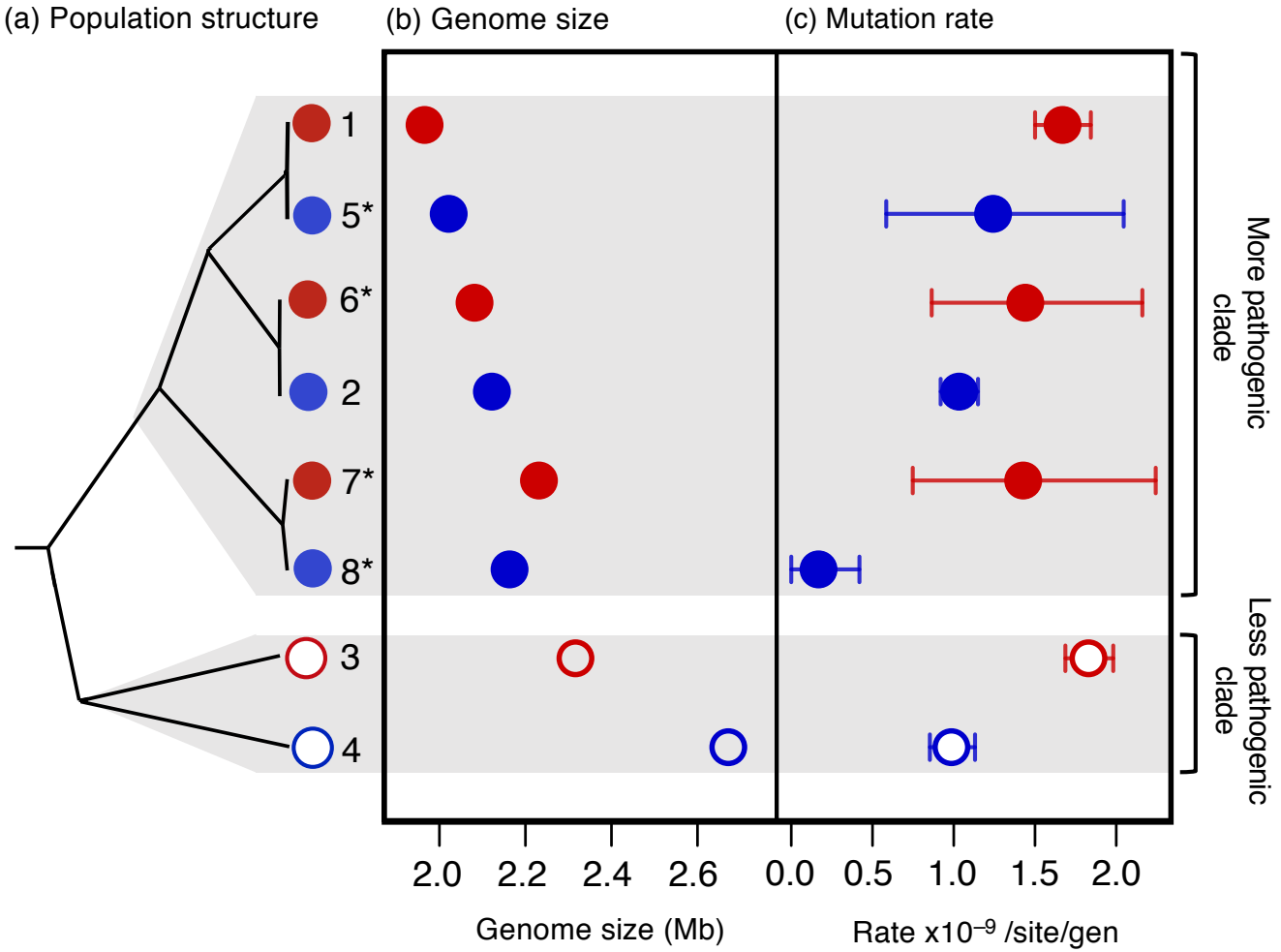
636 **Table 1. Characteristics of the eight strains of *Streptococcus suis* used in both the 200-day (not shaded) and 25-day (shaded) MA**
637 **experiments and estimates of their mutation rates.**
638

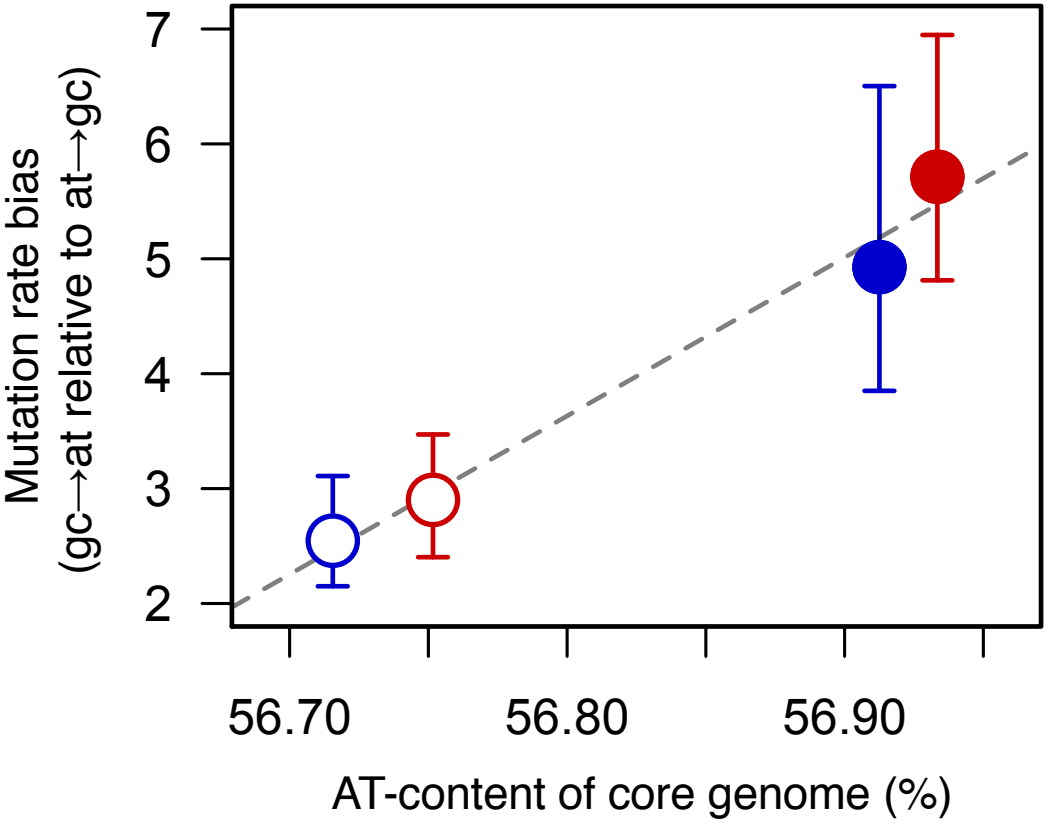
Group	More pathogenic						Less pathogenic	
Strain	1	5	6	2	7	8	3	4
Disease-association	Disease	Carriage	Disease	Carriage	Disease	Carriage	Disease	Carriage
Genome size (Mb)	1.97	2.02	2.08	2.12	2.23	2.16	2.32	2.67
Duration of experiment (days)	200	25	25	200	25	25	200	200
Number of generations	3,991	523	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)

639


(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) Mutation rate
1
5*
6*
2
7*
8*
3
4
More pathogenic clade
Less pathogenic clade
2.0
2.2
2.4
2.6
0.0
0.5
1.0
1.5
2.0
Genome size (Mb)
Rate x10^{-9} /site/gen


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 (%)