Population genomics of *Bacillus anthracis* from an anthrax hyperendemic area reveals transmission processes across spatial scales and unexpected within-host diversity

Taya L. Forde[1*], Tristan P. W. Dennis[1], O. Rhoda Aminu[1], William T. Harvey[1], Ayesha Hassim[2], Ireen Kiwelu[3], Matej Medvecky[1], Deogratius Mshanga[4], Henriette Van Heerden[2], Adeline Vogel[1], Ruth N. Zadoks[1,#], Blandina T. Mmbaga[3,5], Tiziana Lembo[1¶], Roman Biek[1¶]

[1] Institute of Biodiversity, Animal Health & Comparative Medicine, University of Glasgow, Glasgow, United Kingdom

[2] Department of Veterinary Tropical Diseases, Faculty of Veterinary Science, University of Pretoria, Onderstepoort, South Africa

[3] Kilimanjaro Clinical Research Institute, Kilimanjaro Christian Medical Centre, Moshi, Tanzania

[4] Tanzania Veterinary Laboratory Agency, Northern Zone, Arusha, Tanzania

[5] Kilimanjaro Christian Medical University College, Moshi, Tanzania

[#] Current address: Sydney School of Veterinary Science, University of Sydney, Sydney, Australia

* Corresponding author

E-mail: taya.forde@glasgow.ac.uk (TLF)

¶ These authors contributed equally to this work.

Abstract

Genomic sequencing has revolutionized our understanding of bacterial disease epidemiology, but remains underutilized for zoonotic pathogens in remote endemic settings. Anthrax, caused by the spore-forming bacterium *Bacillus anthracis*, remains a threat to human and animal health and rural livelihoods in low- and middle-income countries. While the global genomic diversity of *B. anthracis* has been well-characterized, there is limited information on how its populations are genetically structured at the scale at which transmission occurs, critical for understanding the pathogen's evolution and transmission dynamics. Using a uniquely rich dataset, we quantified genome-wide single nucleotide polymorphisms (SNPs) among 73 *B. anthracis* isolates derived from 33 livestock carcasses sampled over one year throughout the Ngorongoro Conservation Area, Tanzania, an area hyperendemic for anthrax. Genome-wide SNPs distinguished 22 unique *B. anthracis* genotypes within the study area. However, phylogeographic structure was lacking, as identical SNP profiles were found throughout the study area, likely the result of the long and variable periods of spore dormancy and long-distance livestock movements. Significantly, divergent genotypes were obtained from spatio-temporally linked cases and even individual carcasses. The high number of SNPs distinguishing isolates from the same host is unlikely to have arisen during infection, as supported by our simulation models. This points to an unexpectedly wide transmission bottleneck for *B. anthracis*, with an inoculum comprising multiple variants being the norm. Our work highlights that inferring transmission patterns of *B. anthracis* from genomic data will require analytical approaches that account for extended and variable environmental persistence as well as co-infection.

42 # Importance

43 Pathogens transmitted between animals and people affect the health and livelihoods of farmers,
44 particularly in developing countries dependent on livestock. Understanding over what distances
45 these pathogens are transmitted and how they evolve is important to inform control strategies
46 towards reducing disease impacts. Information on the circulation of *Bacillus anthracis*, which
47 causes the often-lethal disease anthrax, is lacking for settings where the disease is commonplace.
48 Consequently, we examined its genetic variability in an area in Tanzania where anthrax is
49 widespread. We found no clear link between how closely cases were sampled and their genetic
50 similarity. We suspect this lack of congruence is primarily driven by large-scale livestock
51 movements, which control efforts should take into consideration. Another significant finding was
52 the co-occurrence of multiple *B. anthracis* types within individual hosts, suggesting animals are
53 commonly infected with a mixture of variants. This needs to be accounted for when investigating
54 possible connections between cases.

55 # Introduction

56 Genomic data have the potential to transform our understanding of the evolution and
57 epidemiology of pathogens of public health importance (1). However, this potential has yet to be
58 fully harnessed for many zoonotic diseases that occur in hard-to-reach areas. Anthrax remains
59 endemic in many low- and middle-income countries (LMICs) worldwide (2). It is a disease
60 characterized by sudden deaths in herbivorous livestock and wildlife, and can also cause serious,
61 potentially fatal disease in people (3). Anthrax is classified among the neglected zoonoses: a
62 group of diseases shared by animals and people that, due to their occurrence in remote,
63 disadvantaged communities, collectively receive less than 0.1% of international global health
64 assistance (4). As for many neglected zoonoses, there is limited genomic data for *Bacillus*
65 *anthracis*, the bacterium that causes anthrax, from endemic LMIC settings where surveillance
66 tends to be limited. Such data could help to improve our understanding of transmission
67 processes, such as how *B. anthracis* is spread within and between outbreaks, and ultimately
68 contribute to more informed disease management.

69 The genomic diversity of *B. anthracis* has been well-described at a global scale. Isolates can be
70 broadly divided into three major clades (A, B, C), of which A clade is the most widespread and
71 globally dominant (5–7). Isolates from most *B. anthracis* lineages have been found across
72 geographically widespread areas, often spanning multiple continents (7). While particular
73 variants often predominate regionally, high lineage diversity has also been reported, including
74 co-circulation of strains from multiple linages (8, 9). How *B. anthracis* diversity is structured at
75 smaller scales is less well-defined. The pathogen has limited genomic diversity compared to
76 other bacterial species (i.e. is genetically monomorphic), rendering standard genotyping methods
77 such as multi-locus sequence typing insufficiently discriminatory (10). A hierarchical genotyping

78 scheme known as PHRANA was therefore developed specifically for *B. anthracis,* based on
79 quickly evolving repetitive regions nested within more phylogenetically stable markers –
80 canSNPs – that distinguish among the major lineages (5, 11). Variants of this scheme have been
81 used to examine the diversity of *B. anthracis* in several endemic settings globally, including in a
82 few African countries (8, 12–14). Genome-wide SNP data would offer higher resolution for
83 discriminating among closely-related isolates. However, whole genome sequencing (WGS)
84 features in only a few studies of local *B. anthracis* diversity (15–18), and has rarely been
85 conducted outside Europe (19), so the potential for phylogenomic data to be used to understand
86 transmission patterns within hyperendemic areas has yet to be explored.

87 Transmission of *B. anthracis* occurs primarily through the environment. After causing the death
88 of the animal host, vegetative bacteria are released into the environment via bodily fluids. Here,
89 upon exposure to oxygen and cues related to a lack of nutrients, these bacteria sporulate and can
90 persist in a dormant yet infectious state for several decades (3, 20). While the viability of spores
91 decreases over time, new *B. anthracis* infections could theoretically arise from recent cases or
92 cases that occurred several years or even decades previously. How this environmental
93 persistence shapes the spatio-temporal diversity of *B. anthracis* in endemic settings has never
94 been investigated.

95 In molecular epidemiological studies of bacterial pathogens, a single isolate is typically
96 sequenced from each individual case. However, this approach fails to recognize the bacterial
97 population diversity that may exist within the host (21, 22). Such diversity can either result from
98 mutations that arise during infection, or from heterogeneity (multiple variants) in the inoculum,
99 either through co-infection (exposure to multiple variants simultaneously) or superinfection
100 (multiple exposures) (23, 24). Under those scenarios, a single isolate is unlikely to represent the

101 overall diversity of the pathogen within the host, and as a result this approach can lead to
102 erroneous inferences about transmission pathways (25). The importance of capturing within-host
103 diversity is therefore increasingly recognized (26, 27). In the case of anthrax, multiple genotypes
104 of *B. anthracis* have been previously found within individual hosts (28–30), but it remains
105 unclear whether this represents a more widespread phenomenon.

106 The objective of this study was to quantify the genomic diversity of *B. anthracis* at hierarchical
107 spatial scales within the livestock population of a hyperendemic setting. This was accomplished
108 using a unique dataset including 1) isolates collected throughout a large (~8,300 km^2) area of
109 northern Tanzania where anthrax is widespread; 2) among spatio-temporally linked cases; and 3)
110 within individual hosts, assessing multiple isolates from the same and different sample types
111 associated with a case (e.g. tissue, blood, soil).

112 Results

113 Genomic sequencing of *B. anthracis* isolates enabled assessment of diversity at 114 hierarchical spatial scales

115 *Bacillus anthracis* isolates were recovered from livestock carcasses in the Ngorongoro
116 Conservation Area (NCA), part of the Serengeti ecosystem of northern Tanzania. The 73 isolates
117 from which WGS data were obtained were from a total of 33 carcass sites sampled throughout
118 the NCA (Table S1). Carcasses were of the following species: sheep (n = 18), cattle (n = 7),
119 goats (n = 4), donkey (n = 1) or unknown host (n = 3). The diversity of *B. anthracis* was assessed
120 at multiple hierarchical spatial scales (Fig. 1). Carcasses for which detailed sampling location
121 data were available (n = 32) could be grouped into four distinct areas within the NCA (central,
122 north, south, east), referred to herein as 'geographical groups'. To assess the genomic relatedness

123 among spatio-temporally-linked cases, on four occasions, samples were collected from two
124 carcasses either from the same or neighboring households on the same or consecutive days,
125 which we refer to as 'epidemiological clusters'. Three of these clusters were in the central
126 geographical group, while the fourth was in the southern group of carcasses sampled. Finally,
127 multiple isolates (n = 2-4) were sequenced from a single carcass for 21 carcasses. These were
128 either i) isolates from multiple sample types (i.e. tissue, blood, swabs, soil and/or insects) from a
129 single carcass (n = 16 carcasses) and/or ii) multiple isolates from the same sample (n = 15
130 carcasses). Isolates contributing to investigations of diversity at each hierarchical level are
131 detailed in Table S1, along with individual sequence quality metrics. Mapped reads had an
132 average depth of coverage of 85X across isolates, ranging from 24X – 245X (median 72X).

133 *B. anthracis* within the NCA is limited to a single subgroup lacking clear 134 phylogeographic signal

135 All *B. anthracis* isolates from the NCA were found to belong to the Ancient A subgroup of Clade
136 A (Fig. 2). Within this subgroup, the NCA isolates formed a monophyletic clade within Cluster
137 3.2 as defined by Bruce et al. (7), which also contained the isolate A2075 (GenBank accession
138 SRR2968187), isolated in 1999 from a baboon in Muhesi Game Reserve, central Tanzania.

139 A total of 125 SNPs polymorphic among the NCA isolates and A2075 were retained for analysis,
140 of which 13 were unique to A2075. Twenty-two unique genotypes (SNP profiles) were found
141 among the 73 NCA isolates (Fig. 3). Based on a rarefaction analysis, the observed genotypic
142 diversity was close to that present throughout the study area (i.e. further sampling would have
143 been unlikely to reveal additional genotypes; Fig. S1). The maximum pairwise nucleotide
144 difference between any two NCA-derived isolates was 49 SNPs [median = 24, interquartile range

145 (IQR) = 10-35] (Fig. S2). There was no clear relationship between the pairwise nucleotide
146 differences and the geographical distance between sampling locations (Fig. S3), as confirmed by
147 a test for isolation by distance (r = 0.04, p = 0.214). Isolates from the central, eastern and
148 southern geographical groups were observed throughout the phylogenetic tree, while all isolates
149 from the northern sampling area were restricted to a single clade that contained the majority of
150 NCA isolates (Fig. 3). In eight instances, identical *B. anthracis* genotypes were found in
151 carcasses from different geographical groups, all of which involved isolates from the central
152 geographical group and one of the other areas, with all areas implicated (Table S3). Identical or
153 nearly identical SNP profiles (1 SNP difference) were obtained from carcasses sampled 3-5
154 months apart on six occasions, and 10 months apart on one occasion (Table S4).

155 *B. anthracis* isolates from spatio-temporally linked cases are rarely genetically 156 related

157 Between one and four isolates were sequenced from each carcass sampled as part of an
158 epidemiological cluster (i.e. pair of spatio-temporally linked carcasses), resulting in between 4-8
159 isolates per cluster. In most cases, isolates deriving from the same epidemiological cluster were
160 phylogenetically unrelated. Only in one of the four clusters examined (C2) did both carcasses
161 contain isolates with identical core SNP profiles (Fig. 3B, S6 and S7). Overall, isolates from
162 different carcasses within the same epidemiological cluster had similar numbers of SNP
163 differences when compared with randomly selected carcasses (linked cases: median = 21, IQR =
164 0-33; unlinked cases: median = 23, IQR = 10-35) (Fig. 4).

165 Within-host diversity of *B. anthracis* is similar to between-host diversity in the 166 NCA

167 Overall, the number of pairwise SNP differences between isolates from the same carcass was
168 lower than that found between isolates from different carcasses [same: median = 11, interquartile
169 range = 0-34; different: median = 23, interquartile range = 10-35] (Fig. 4). However, isolates
170 with multiple distinct genotypes were obtained on 15 of 21 occasions wherein multiple isolates
171 were sequenced from the same carcass, with as many as 43 SNP differences between isolates
172 (Fig. 5). A high level of divergence was seen, regardless of whether isolates were from the same
173 or different sample type (e.g. multiple isolates from a tissue sample vs. isolates from tissue and
174 soil samples, Fig. 5A). Divergent genotypes were observed in all carcasses from which three or
175 more isolates were sequenced (12/12), and in a third (3/9) of carcasses for which two isolates
176 were sequenced. A single SNP difference separated one isolate from two others within one
177 carcass (AN16-83) and two SNP differences separated two isolates from a single soil sample
178 (LNA). All other pairwise within-host SNP differences were eight or above. All of the within-
179 host SNP profiles detected were shared with those from other carcasses.

180 Observed within-host diversity is unlikely to have arisen during the course of 181 infection

182 To assess the likelihood of different levels of within-host diversity arising during the course of
183 infection, we performed simulation modelling of *B. anthracis* infection with a homogenous
184 inoculum of varying size (1, 2, 5, 10, 20, 50 and 100 bacterial genomes), and across a range of
185 replication cycles (up to 25 generations) using previously established mutation rates (29). With
186 an infectious dose d, discrete generations and no die-off, the population size in generation i is
187 therefore $d \times 2^{(i-1)}$. Averaging across simulations with varying initial doses, in the 25th
188 generation, 90.0% of the population was identical to the infecting dose, 9.53% differed by 1
189 SNP, and 0.489% by 2 SNPs, with higher numbers of mutations very rare (<0.01%). While

190 stochasticity was present in early generations and particularly at low inoculum doses (Fig. S8-
191 10), the mean proportion of genomes with various numbers of SNPs changed predictably after
192 the initial generations (Fig. 6A-B). Extrapolating to the 40th generation, then ~84% of the
193 population would be expected to be identical to the infecting dose, ~15% differing by 1 SNP and
194 ~0.9% by 2 SNPs.

195 To illustrate how this within-host generated diversity would be captured in our sampling, pairs of
196 genomes were repeatedly sampled from simulated populations and pairwise differences
197 calculated (Fig 6C and 6D, File S1). Overall, pairwise differences greater than 2 SNPs occurred
198 in less than 0.2% of all simulated populations (Table S5). Higher SNP distances such as those
199 observed (Fig. 5B) are therefore unlikely to arise within-host following infection with a
200 homogenous inoculum.

201 Discussion

202 Genomic data for understanding the population structure and transmission patterns of bacterial
203 zoonoses has been limited for LMIC settings where these diseases tend to have the greatest
204 impact. Despite *B. anthracis* having limited genomic diversity in comparison with other bacterial
205 species, WGS provided sufficient resolution to discriminate among isolates collected from within
206 a relatively small geographic area of a few thousand square kilometers. The way in which the
207 genomic diversity was partitioned across hierarchical spatial scales within this area has a series
208 of novel implications for our understanding of how the pathogen is transmitted and evolves
209 during endemic circulation.

210 *B. anthracis* diversity within the hyperendemic NCA region is limited to a
211 single Clade A sub-group

212 The NCA is an area where anthrax has likely been endemic for decades, if not centuries. Local
213 community members claim that it has been an issue for their health and that of their livestock
214 throughout living memory. Despite WGS providing sufficient discriminatory power to
215 differentiate among individual *B. anthracis* isolates in this setting, fewer than 50 SNP differences
216 were found among NCA isolates across the 5.2 MB chromosomal genome, highlighting the
217 degree to which this bacterial pathogen is monomorphic. All 73 sequenced isolates formed a
218 monophyletic cluster within the Ancient A lineage – also known as canSNP group A.Br.005/006
219 or A.Br.034 (6) – a subgroup of Clade A comprised mostly of isolates from south-eastern Africa.
220 This contrasts with some previous studies of anthrax diversity in chronically endemic areas
221 reporting co-circulation of isolates from multiple lineages using lower resolution markers (8, 9).
222 Our study, which only included isolates collected over one year, can be considered a snapshot of
223 recent diversity only. That being said, Tanzanian isolates genotyped in a previous study (n = 17)
224 were also all found to belong to Ancient A clade (5), demonstrating that this particular lineage is
225 dominant and well-established in this country. The most closely related publicly available isolate
226 (A2075), sampled in central Tanzania approximately 300km away from our study area, differed
227 from some NCA isolates by only 13 SNPs, which illustrates that highly related *B. anthracis*
228 isolates can be geographically widespread. Broader, longitudinal WGS studies of *B. anthracis*
229 across different regions of Africa will be needed to assess the actual range of individual
230 genotypes over space and time.

231 No phylogeographic signal observed despite considerable SNP diversity

232 Despite the considerable diversity observed (i.e. 22 unique genotypes within the sampled
233 population), a phylogeographic signal was not detected at the scale of this study area. The
234 finding of identical SNP profiles across distances of tens of kilometers likely reflects the ecology

235 of anthrax in general and in our study system. First, there are few opportunities for genetic
236 diversity to arise within the *B. anthracis* lifecycle, which is characterized by long periods of
237 environmental dormancy in spore form, punctuated by brief interludes of a few days where it
238 develops into its vegetative state and replicates within an infected host. It is estimated that *B.*
239 *anthracis* undergoes only 20-40 replications per infection (11). Given its low mutation rate (5.2 –
240 8.3 x 10^{-10} mutations/site/generation) (5, 29, 31), novel mutations would likely arise in only a
241 small proportion of *B. anthracis* infections, as also supported by our simulations. The short time
242 of active replication within a host also means there is minimal opportunity for horizontal gene
243 transfer with other bacteria (11), which further restricts the ability of *B. anthracis* to diversify. It
244 is believed that *B. anthracis* rarely multiplies outside of a host, although there is some evidence
245 for limited environmental replication (32, 33). Viability of *B. anthracis* decays exponentially
246 over time, although infectious spores remain detectable at carcass sites at least four years after
247 the death of the animal (34), and under favorable conditions, spores can remain viable for up to
248 several decades (20). It is therefore reasonable to expect that spores from a single anthrax carcass
249 with few or no novel mutations could be the source of subsequent infections over highly variable
250 time periods, a phenomenon referred to by Sahl et al. (2016) as a 'time capsule'. This situation
251 violates typical molecular clock assumptions (35), wherein SNPs would be expected to arise at a
252 relatively steady rate within the pathogen population over time. The observation that
253 contemporary isolates from the NCA are phylogenetically basal to isolate A2075, collected
254 nearly two decades earlier, highlights this issue. In order for molecular clock models to be
255 applied to environmentally-persistent pathogens such as *B. anthracis* – for instance to estimate
256 how long particular lineages have been in circulation – current analytical frameworks will need
257 to be extended.

258 In addition to long and variable environmental persistence, animal movements and spatial
259 admixture likely contribute to the lack of phylogeographic structure of *B. anthracis* within this
260 study area. Potential sources of infection are quite spatially restricted, as the highest
261 concentrations of viable spores are found within only a few meters of anthrax carcasses (36), in
262 what have been termed 'localized infectious zones' (37). However, given that the incubation
263 period of anthrax in livestock typically ranges from 1-14 days and potentially longer (3),
264 extensive movement can occur between the time of infection and the animal's death, meaning
265 the site of sampling likely does not reflect the site of infection. In parts of rural Africa where
266 pastoralism is the main form of agriculture, livestock are moved for various reasons, including to
267 access water, pastures and minerals. Such daily and/or seasonal livestock movements could
268 contribute to the observation of identical SNP profiles over distances of tens of kilometers in our
269 study area and the lack of relationship between genetic distance and geographic distance between
270 sampling locations. While livestock appear to be the primary drivers of *B. anthracis* transmission
271 in the NCA, movement of infected wildlife and scavengers acting as carriers of *B. anthracis*
272 spores could represent an additional mechanism contributing to our observations (38, 39).
273 Further comparative genomic studies across wider areas will be essential for elucidating the
274 geographic scales at which transmission occurs. This would help to delineate areas across which
275 coordinated livestock vaccination campaigns should occur to avoid regular re-incursions through
276 animal movement and trade.

277 High within-host diversity is the result of simultaneous infection with multiple 278 variants, not within-host evolution

279 We observed high *B. anthracis* diversity within individual hosts, essentially indistinguishable
280 from levels of diversity found throughout the study area. Smaller numbers of SNPs (1-2) could

281 potentially have arisen during culture; however, based on previous passaging experiments with
282 *B. anthracis* (29), we do not expect this to have contributed in our case due to the limited number
283 of passages performed. Alternatively, small numbers of SNPs could have arisen during the
284 course of infection within the host. The simulations we performed suggest that isolates with >2
285 SNP differences between them are unlikely to be the product of within-host evolution. The
286 regular occurrence of such differences or greater among isolates from the same carcass indicates
287 that animals are commonly infected with a heterogenous infectious dose (i.e. ingestion of a
288 mixture of *B. anthracis* genotypes from single or possibly multiple grazing or watering points;
289 Fig. 7). Multiple SNP profiles were present among isolates from the same individual soil
290 samples, supporting the occurrence of heterogeneity in a single infection source. It is also
291 noteworthy that all within-host SNP profiles observed in this study were shared with isolates
292 from at least one other sampled carcass (i.e. none were unique); this strongly suggests that most
293 of the observed within-host diversity is the result of various genotypes having been present in the
294 inoculum rather than having been generated *de novo*. This points to a wide transmission
295 bottleneck, since a small inoculum comprising only a few spores would limit the possible
296 diversity that could be transmitted.

297 Our findings make an important contribution to the ongoing debate about the size of the
298 transmission bottleneck for naturally-occurring anthrax in animals (i.e. the number of spores that
299 give rise to a case). Recent work has proposed that founding populations may be as small as 1-3
300 individual spores (30). However, our findings are clearly at odds with such narrow bottlenecks,
301 since the multiple genotypes we regularly observed within individual hosts could not have been
302 transmitted within such a small inoculum. While our results provide no information about the
303 exact size of the infectious dose, they align more closely with earlier suggestions of higher

304 infectious doses, which are biologically plausible given that animals grazing at carcass sites
305 might ingest hundreds of thousands of spores with each bite (34).

306 Spatio-temporally linked anthrax cases are rarely genetically linked

307 Limiting phylogenetic analyses to a single isolate per host often leads to incorrect inference of
308 transmission events (22), including the potential to overlook important epidemiological
309 connections (40). This issue is exemplified in the current study by the spatio-temporally linked
310 pair of cases in Cluster 2: whereas several of the isolates from both carcasses (n = 3 each) had
311 identical SNP profiles, supporting a transmission link, both carcasses also harbored non-identical
312 genotypes (Fig. S5 and S6). Under these circumstances, transmission links may be missed, even
313 with complete sampling. As noted by Ågren et al. (2014), in the case of multi-clonal infections,
314 subsequent cases may stem from different genotypes from within the founding population,
315 masking the fact that these cases stemmed from a common source, regardless of the number of
316 isolates sequenced among the subsequent cases. This could be the case for the other three spatio-
317 temporally linked pairs of anthrax carcasses investigated in this study (Fig. S6), in which we
318 only detected isolates with distinct SNP profiles. Alternatively, cases without a genetic link
319 could be temporally linked for reasons other than exposure to a common source. For instance,
320 animals might be more susceptible or at greater risk of exposure to infection at particular times
321 of year, e.g. due to lower immune function related to their nutritional status and/or associated
322 with weather extremes including prolonged rains or droughts (8). More extensive sampling of
323 within-carcass diversity would be necessary to investigate these hypotheses and to determine
324 whether co-occurrence of genotypes could be used to track transmission patterns.

Conclusions

In this study, the genomic diversity of *B. anthracis* was quantified at various spatial scales within a hyperendemic setting. While WGS could discriminate among isolates within a relatively small geographic area, there was a lack of phylogeographic signal and limited genetic relatedness was observed among isolates from spatio-temporally linked cases. We hypothesize that this lack of spatial structure reflects the long-term persistence of *B. anthracis* spores in the environment, combined with extensive livestock movements related to local pastoralist practices. Based on simulations, the high within-host heterogeneity we observed points to an inoculum comprised of diverse genotypes, suggestive of a wide transmission bottleneck. Our work paves the way for studying *B. anthracis* genomic diversity and evolution within anthrax-endemic areas more broadly and to confirm the temporal and spatial scales over which genomic data are most informative for inferring transmission dynamics.

Methods

Study area

This study was conducted in the NCA of northern Tanzania, which covers 8,292 km^2. Located to the south-east of Serengeti National Park, this multiple-land use area is inhabited by roughly 87,000 people (41) and one million livestock (sheep, goats and cattle) (Veterinary Officer for Ngorongoro District, pers comm). Northern Tanzania remains hyperendemic for anthrax (42), and prior to this study the NCA was recognized as a potential hotspot for this disease (43).

Research and ethical approval

345 This study received ethical approval from the Kilimanjaro Christian Medical University College
346 Ethics Review Committee (certificate No. 2050); the National Institute for Medical Research,
347 Tanzania (NIMR/HQ/R.8a/Vol. IX/2660); Tanzanian Commission for Science and Technology
348 (2016-95-NA-2016-45); and the College of Medical Veterinary and Life Sciences ethics
349 committee at the University of Glasgow (200150152). It also received permission under Section
350 20 of the Animal Diseases Act 35 (1984) at the University of Pretoria, South Africa (Ref
351 12/11/1/1/6).

352 Sample collection

353 Samples were collected between May 2016 and April 2017 inclusive through active surveillance
354 by a dedicated field team. Sudden deaths in animals reported by community members throughout
355 the NCA were investigated and samples were collected when anthrax was suspected (File S1).
356 When available, the following samples were collected: a piece of skin tissue (tip of the ear if the
357 carcass was still intact, or a piece of hide if the carcass had already been opened); whole blood;
358 swab of blood or body fluid at natural orifices; blood- or body fluid-soaked soil from below the
359 carcass; and insects found on or around the carcass. Various metadata were recorded, including
360 the species of animal affected and the location of sampling (Table S1, File S1). All samples were
361 stored at ambient temperature for up to six months at local veterinary facilities until transport to
362 the Kilimanjaro Clinical Research Institute (KCRI) in Moshi, Tanzania for molecular
363 diagnostics, as previously described (44), with aliquots shipped to the University of Pretoria,
364 South Africa, for selective culture and DNA extraction from *B. anthracis* isolates.

365 Selective culture, DNA extraction and sequencing

366 Sample pretreatment (i.e. to inhibit competition from heat-sensitive bacteria) is described in File
367 S1. Sample homogenates (100 µL) were plated onto both polymyxin-EDTA thallous acetate
368 (PET) selective media and 5% sheep blood agar (SBA). These were incubated at 37 °C overnight
369 and the plates inspected for growth after 15 – 24 hours incubation. The PET was then further
370 incubated and inspected at 48 hours. Suspect *B. anthracis* colonies based on typical
371 morphological characteristics were sub-cultured onto SBA for purification and identification
372 (File S1). In parallel, a single colony was streaked onto a new purity plate for nucleic acid
373 extraction. In some instances multiple isolates were selected from the same sample where the
374 colonies demonstrated differences in morphology but were identified on the same plate and met
375 the selection criteria (File S1).

376 DNA extracts from 75 isolates from 33 carcasses were submitted for library preparation and
377 sequencing at MicrobesNG (Birmingham, UK). Libraries were prepared using the Nextera XT
378 v2 kit (Illumina, San Diego, USA) and sequenced on the Illumina HiSeq platform, generating
379 250 base pair paired-end reads.

380 Bioinformatics and genomic analyses

381 Reads were adapter trimmed by MicrobesNG using Trimmomatic v0.30 (45), and basic statistics
382 determined using QUAST (46) (Table S1). Bacterial species identification was confirmed using
383 Kraken (47). Based on these quality metrics, sequences from two isolates were excluded from
384 further analyses: one due to a low number of reads (<40,000), and another that was identified as
385 *B. cereus*. There were some further indications that not all cultures were pure *B. anthracis,*
386 despite multiple rounds of sub-culture (File S1, Table S1). A reference-based mapping approach
387 and strict variant filtering criteria were implemented to minimize the issues associated with the
388 sequence quality while making use of as much of the data as possible.

389 Read mapping and variant calling were performed on the CLIMB computing platform for
390 microbial genomics (48). Trimmed reads were aligned to the chromosome of the Ames Ancestor
391 reference genome (NC_007530) using bwa-mem (version 0.7.17). Picard was used to mark and
392 remove duplicate reads, add read group information, and index the bam files (49). Quality
393 metrics for read mapping were obtained using Qualimap (50) (Table S1). SNPs were detected in
394 individual isolates by VarScan v2.4.4 (51) with parameters set as follows: minimum read depth
395 of 4; minimum base quality of 20; variant allele frequency ≥ 0.95. Subsequent SNP curation
396 steps are described in File S1. Custom python scripts for the assessment of read mapping SNP
397 metrics data, variant site filtering and generation of variant call and alignment files (source codes
398 with description of their functionality and usage), along with the final variant call and multiple
399 sequence alignment files are available on GitHub
400 (https://github.com/matejmedvecky/anthraxdiversityscripts).

401 The alignment of concatenated SNPs was analyzed using ModelFinder to determine the most
402 appropriate model of nucleotide substitution (52). Subsequently, a maximum-likelihood
403 phylogeny was estimated in IQ-Tree (53) under the Kimura-3-parameters (K3P) model, using
404 1000 ultrafast bootstrap replicates (54). A distance matrix detailing SNP differences between
405 isolates was constructed using snp-dists v0.6 (https://github.com/tseemann/snp-dists). The
406 distance between GPS points was calculated using the pointDistance command in the R package
407 *raster* (55). Isolation by distance was tested using Mantel test to assess the correlation between
408 SNP distance and Euclidean geographic distance within the R package *adegenet* (56). All
409 program versions and commands used, the distance matrix, as well as small custom scripts are
410 available on GitHub (https://github.com/tristanpwdennis/anthrax_diversity).

411 To place the newly sequenced isolates within the global phylogeny of *B. anthracis*, 80 WGS
412 from GenBank were accessed (Table S2), and a core genome alignment generated using Parsnp
413 v1.1.2 (57). The resulting phylogeny indicated that NCA isolates belong to the Ancient A
414 lineage. To further resolve the diversity among NCA isolates compared with other publicly
415 available isolates from the same lineage, reads from eight additional isolates available on SRA
416 from a study by Bruce et al. 2019 were accessed using fastq-dump from the SRA-toolkit: all
417 isolates (n=4) belonging to the 3.2 linage, and two arbitrarily selected isolates from each the 3.1
418 and 3.3 lineages (Table S2). These were run through our SNP-calling pipeline as described
419 above, resulting in a sequence alignment file free of -/N characters, which was used to infer a
420 phylogeny in RAxML v8.2.11 (58) using a GTR model of nucleotide substitution, and using the
421 Ames Ancestor reference genome as an outgroup.

422 Simulation modelling

423 The initial number of genomes in the inoculum was varied between one and 100 genomes to
424 reflect uncertainty in anthrax infectious dose. In each bacterial generation, genomes underwent a
425 round of replication followed by cell division and accordingly population size doubled in each
426 generation. The number of mutations occurring in the replication of each genome was drawn
427 from a Poisson distribution parameterized to reflect the estimated *B. anthracis* mutation rate (i.e.
428 $\lambda = 0.004316$). This genome-level mutation rate is based on the genome size of 5.2 million base
429 pairs and a mutation rate of 8.3×10^{-10} mutations per site (29); this represents an upper estimate
430 of the mutation rate and was chosen as we were interested in estimating the upper limits on
431 reasonable expectation of diversity emerging during the course of an infection initiated by a
432 homogenous dose. Simulations were carried out for up to 30 generations (replication cycles), and
433 results extrapolated to 40 generations, proposed to be the upper limit on the number of

434 replication cycles during an infection (11). Details of simulations run are provided in File S1.
435 Pairs of genomes were repeatedly sampled from these simulated populations and the count of
436 mutations separating each pair was calculated. Sampling was performed 100 times per
437 generation for each simulation. Simulations, sampling of simulated populations, and linear
438 models summarizing trends in the outcome of these processes were performed in R (59).

439 Data availability

440 Raw sequencing reads are available on European Nucleotide Archive SRA under accession
441 number PRJEB45684.

442 Acknowledgements

443 We are grateful for all the support received for this research, particularly from the NCA
444 community and authorities. We thank the Ngorongoro District Council, Ngorongoro
445 Conservation Area Authority, District Veterinary Office, Tanzania Wildlife Research Institute
446 and members of our field team – Sabore Ole Moko, Sironga Nanjicho, Kadogo Lerimba and
447 Godwin Mshumba – for assistance with this study. We also thank the Directorate of Veterinary
448 Services, Ministry of Agriculture, Livestock and Fisheries, and Ministry of Health, Community
449 Development, Gender, Elderly and Children for their support. Finally, we thank Yi Xuan Chew
450 and Nichith Kollanandi Ratheesh for initial assessments of the publicly available genomic data.
451 Establishment and maintenance of the Zoonoses laboratory at KCRI (BTM) was supported by
452 the BBSRC (BB/J010367/1), BBSRC Zoonoses and Emerging Livestock Systems (BB/L017679,
453 BB/L018926, BB/L018845), and the Wellcome Trust (096400/Z/11/Z).

454 # References

455 1. Kao RR, Haydon DT, Lycett SJ, Murcia PR. 2014. Supersize me: how whole-genome
456 sequencing and big data are transforming epidemiology. Trends Microbiol 22:282–91.

457 2. Carlson CJ, Kracalik IT, Ross N, Alexander KA, Hugh-Jones ME, Fegan M, Elkin BT, Epp
458 T, Shury TK, Zhang W, Bagirova M, Getz WM, Blackburn JK. 2019. The global
459 distribution of *Bacillus anthracis* and associated anthrax risk to humans, livestock and
460 wildlife. Nat Microbiol 4:1337–1343.

461 3. WHO. 2008. Anthrax in Humans and Animals. 4th Ed.

462 4. WHO. 2011. The control of neglected zoonotic diseases. Community-based interventions
463 for prevention and control. Report of the third conference organized with ICONZ, DFID-
464 RIU, Gates Foundation, SOS, EU, TDR and FAO with the participation of ILRI and OIE.,
465 Geneva, Switzerland.

466 5. Van Ert MN, Easterday WR, Huynh LY, Okinaka RT, Hugh-Jones ME, Ravel J, Zanecki
467 SR, Pearson T, Simonson TS, U'Ren JM, Kachur SM, Leadem-Dougherty RR, Rhoton SD,
468 Zinser G, Farlow J, Coker PR, Smith KL, Wang B, Kenefic LJ, Fraser-Liggett CM, Wagner
469 DM, Keim P. 2007. Global genetic population structure of *Bacillus anthracis*. PLoS ONE
470 2:e461.

471 6. Sahl JW, Pearson T, Okinaka R, Schupp JM, Gillece JD, Heaton H, Birdsell D, Hepp C,
472 Fofanov V, Noseda R, Fasanella A, Hoffmaster A, Wagner DM, Keim P. 2016. A *Bacillus*
473 *anthracis* genome sequence from the Sverdlovsk 1979 autopsy specimens. mBio 7: e01501-
474 16.

475 7. Bruce SA, Schiraldi NJ, Kamath PL, Easterday WR, Turner WC. 2020. A classification
476 framework for *Bacillus anthracis* defined by global genomic structure. Evol Appl 13:935–
477 944.

478 8. Smith KL, DeVos V, Bryden H, Price LB, Hugh-Jones ME, Keim P. 2000. *Bacillus*
479 *anthracis* diversity in Kruger National Park. J Clin Microbiol 38:3780–3784.

480 9. Simonson TS, Okinaka RT, Wang B, Easterday WR, Huynh L, U'Ren JM, Dukerich M,
481 Zanecki SR, Kenefic LJ, Beaudry J, Schupp JM, Pearson T, Wagner DM, Hoffmaster A,
482 Ravel J, Keim P. 2009. *Bacillus anthracis* in China and its relationship to worldwide
483 lineages. BMC Microbiol 9:71.

484 10. Achtman M. 2008. Evolution, population structure, and phylogeography of genetically
485 monomorphic bacterial pathogens. Annu Rev Microbiol 62:53–70.

486 11. Keim P, Van Ert MN, Pearson T, Vogler AJ, Huynh LY, Wagner DM. 2004. Anthrax
487 molecular epidemiology and forensics: using the appropriate marker for different
488 evolutionary scales. Infect Genet Evol 4:205–213.

489 12. Fasanella A, Serrecchia L, Chiaverini A, Garofolo G, Muuka GM, Mwambazi L. 2018. Use
490 of Canonical Single Nucleotide Polymorphism (CanSNPs) to characterize *Bacillus*
491 *anthracis* outbreak strains in Zambia between 1990 and 2014. PeerJ 6:e5270.

492 13. Beyer W, Bellan S, Eberle G, Ganz HH, Getz WM, Haumacher R, Hilss KA, Kilian W,
493 Lazak J, Turner WC, Turnbull PCB. 2012. Distribution and molecular evolution of *Bacillus*
494 *anthracis* genotypes in Namibia. PLoS Negl Trop Dis 6:e1534.

495 14. Maho A, Rossano A, Hächler H, Holzer A, Schelling E, Zinsstag J, Hassane MH,
496 Toguebaye BS, Akakpo AJ, Van Ert M, Keim P, Kenefic L, Frey J, Perreten V. 2006.
497 Antibiotic susceptibility and molecular diversity of *Bacillus anthracis* strains in Chad:
498 detection of a new phylogenetic subgroup. J Clin Microbiol 44:3422–3425.

499 15. Girault G, Blouin Y, Vergnaud G, Derzelle S. 2014. High-throughput sequencing of
500 *Bacillus anthracis* in France: investigating genome diversity and population structure using
501 whole-genome SNP discovery. BMC Genomics 15:288.

502 16. Derzelle S, Girault G, Roest HIJ, Koene M. 2015. Molecular diversity of *Bacillus anthracis*
503 in the Netherlands: investigating the relationship to the worldwide population using whole-
504 genome SNP discovery. Infect Genet Evol 32:370–376.

505 17. Derzelle S, Girault G, Kokotovic B, Angen Ø. 2015. Whole genome-sequencing and
506 phylogenetic analysis of a historical collection of *Bacillus anthracis* strains from Danish
507 cattle. PLoS ONE 10:e0134699.

508 18. Lienemann T, Beyer W, Pelkola K, Rossow H, Rehn A, Antwerpen M, Grass G. 2018.
509 Genotyping and phylogenetic placement of *Bacillus anthracis* isolates from Finland, a
510 country with rare anthrax cases. BMC Microbiology 18:102.

511 19. Lekota KE, Hassim A, Madoroba E, Hefer CA, van Heerden H. 2020. Phylogenomic
512 structure of *Bacillus anthracis* isolates in the Northern Cape Province, South Africa
513 revealed novel single nucleotide polymorphisms. Infection, Genetics and Evolution
514 80:104146.

515 20. Carlson CJ, Getz WM, Kausrud KL, Cizauskas CA, Blackburn JK, Bustos Carrillo FA,
516 Colwell R, Easterday WR, Ganz HH, Kamath PL, Økstad OA, Turner WC, Kolstø A-B,
517 Stenseth NC. 2018. Spores and soil from six sides: interdisciplinarity and the environmental
518 biology of anthrax (*Bacillus anthracis*). Biol Rev Camb Philos Soc 93:1813–1831.

519 21. Döpfer D, Buist W, Soyer Y, Munoz MA, Zadoks RN, Geue L, Engel B. 2008. Assessing
520 genetic heterogeneity within bacterial species isolated from gastrointestinal and
521 environmental samples: how many isolates does it take? Appl Environ Microbiol 74:3490–
522 3496.

523 22. Worby CJ, Lipsitch M, Hanage WP. 2014. Within-host bacterial diversity hinders accurate
524 reconstruction of transmission networks from genomic distance data. PLoS Comput Biol
525 10:e1003549.

526 23. Futse JE, Brayton KA, Dark MJ, Knowles DP, Palmer GH. 2008. Superinfection as a driver
527 of genomic diversification in antigenically variant pathogens. Proc Natl Acad Sci USA
528 105:2123–2127.

529 24. Sintchenko V, Holmes EC. 2015. The role of pathogen genomics in assessing disease
530 transmission. BMJ 350:h1314.

531 25. Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. 2016. Within-host evolution of
532 bacterial pathogens. Nat Rev Microbiol 14:150–162.

533 26. Paterson GK, Harrison EM, Murray GGR, Welch JJ, Warland JH, Holden MTG, Morgan
534 FJE, Ba X, Koop G, Harris SR, Maskell DJ, Peacock SJ, Herrtage ME, Parkhill J, Holmes

535 MA. 2015. Capturing the cloud of diversity reveals complexity and heterogeneity of MRSA
536 carriage, infection and transmission. Nat Commun 6:6560.

537 27. Alamil M., Hughes J., Berthier K., Desbiez C., Thébaud G., Soubeyrand S. 2019. Inferring
538 epidemiological links from deep sequencing data: a statistical learning approach for human,
539 animal and plant diseases. Philosophical Transactions of the Royal Society B: Biological
540 Sciences 374:20180258.

541 28. Beyer W, Turnbull PCB. 2013. Co-infection of an animal with more than one genotype can
542 occur in anthrax. Lett Appl Microbiol 57:380–384.

543 29. Ågren J, Finn M, Bengtsson B, Segerman B. 2014. Microevolution during an anthrax
544 outbreak leading to clonal heterogeneity and penicillin resistance. PLoS ONE 9:e89112.

545 30. Easterday WR, Ponciano JM, Gomez JP, Van Ert MN, Hadfield T, Bagamian K, Blackburn
546 JK, Stenseth NC, Turner WC. 2020. Coalescence modeling of intrainfection *Bacillus*
547 *anthracis* populations allows estimation of infection parameters in wild populations. Proc
548 Natl Acad Sci USA 117:4273-4280.

549 31. Vogler AJ, Busch JD, Percy-Fine S, Tipton-Hunton C, Smith KL, Keim P. 2002. Molecular
550 analysis of rifampin resistance in *Bacillus anthracis* and *Bacillus cereus*. Antimicrob
551 Agents Chemother 46:511–513.

552 32. Saile E, Koehler TM. 2006. *Bacillus anthracis* multiplication, persistence, and genetic
553 exchange in the rhizosphere of grass plants. Appl Environ Microbiol 72:3168–3174.

554 33. Braun P, Grass G, Aceti A, Serrecchia L, Affuso A, Marino L, Grimaldi S, Pagano S,
555 Hanczaruk M, Georgi E, Northoff B, Schöler A, Schloter M, Antwerpen M, Fasanella A.
556 2015. Microevolution of Anthrax from a Young Ancestor (M.A.Y.A.) suggests a soil-borne
557 life cycle of *Bacillus anthracis*. PLoS ONE 10:e0135346.

558 34. Turner WC, Kausrud KL, Beyer W, Easterday WR, Barandongo ZR, Blaschke E, Cloete
559 CC, Lazak J, Van Ert MN, Ganz HH, Turnbull PCB, Stenseth NC, Getz WM. 2016. Lethal
560 exposure: An integrated approach to pathogen transmission via environmental reservoirs.
561 Sci Rep 6:27311.

562 35. Biek R, Pybus OG, Lloyd-Smith JO, Didelot X. 2015. Measurably evolving pathogens in
563 the genomic era. Trends in Ecology & Evolution 30:306–313.

564 36. Dragon DC, Bader DE, Mitchell J, Woollen N. 2005. Natural dissemination of *Bacillus*
565 *anthracis* spores in northern Canada. Appl Environ Microbiol 71:1610–1615.

566 37. Blackburn JK, Ganz HH, Ponciano JM, Turner WC, Ryan SJ, Kamath P, Cizauskas C,
567 Kausrud K, Holt RD, Stenseth NC, Getz WM. 2019. Modeling R_0 for pathogens with
568 environmental transmission: Animal movements, pathogen populations, and local infectious
569 zones. Int J Environ Res Public Health 16:954.

570 38. 1979. Vultures as carriers of anthrax. J S Afr Vet Assoc 50:35.

571 39. Stears K, Schmitt MH, Turner WC, McCauley DJ, Muse EA, Kiwango H, Mathayo D,
572 Mutayoba BM. 2021. Hippopotamus movements structure the spatiotemporal dynamics of
573 an active anthrax outbreak. Ecosphere 12:e03540.

574 40. Forde TL, Orsel K, Zadoks RN, Biek R, Adams LG, Checkley SL, Davison T, De Buck J,
575 Dumond M, Elkin BT, Finnegan L, Macbeth BJ, Nelson C, Niptanatiak A, Sather S,
576 Schwantje HM, Van Der Meer F, Kutz SJ. 2016. Bacterial genomics reveal the complex
577 epidemiology of an emerging pathogen in arctic and boreal ungulates. Front Microbiol
578 7:1759.

579 41. National Bureau of Statistics, Office of Chief Government Statistician. 2013. 2012
580 Population and Housing Census. The United Republic of Tanzania.

581 42. Mwakapeje ER, Høgset S, Fyumagwa R, Nonga HE, Mdegela RH, Skjerve E. 2018.
582 Anthrax outbreaks in the humans - livestock and wildlife interface areas of Northern
583 Tanzania: a retrospective record review 2006–2016. BMC Public Health 18:106.

584 43. Lembo T, Hampson K, Auty H, Beesley CA, Bessell P, Packer C, Halliday J, Fyumagwa R,
585 Hoare R, Ernest E, Mentzel C, Mlengeya T, Stamey K, Wilkins PP, Cleaveland S. 2011.
586 Serologic surveillance of anthrax in the Serengeti ecosystem, Tanzania, 1996-2009.
587 Emerging Infect Dis 17:387–394.

588 44. Aminu OR, Lembo T, Zadoks RN, Biek R, Lewis S, Kiwelu I, Mmbaga BT, Mshanga D,
589 Shirima G, Denwood M, Forde TL. 2020. Practical and effective diagnosis of animal
590 anthrax in endemic low-resource settings. PLOS Neglected Tropical Diseases 14:e0008655.

591 45. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina
592 sequence data. Bioinformatics 30:2114–2120.

593 46. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for
594 genome assemblies. Bioinformatics 29:1072–1075.

595 47. Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using
596 exact alignments. Genome Biol 15:R46.

597 48. Connor TR, Loman NJ, Thompson S, Smith A, Southgate J, Poplawski R, Bull MJ,
598 Richardson E, Ismail M, Elwood-Thompson S, Kitchen C, Guest M, Bakke M, Sheppard
599 SK, Pallen MJ. 2016. CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an
600 online resource for the medical microbiology community. Microb Genom 2:e000086.

601 49. Broad Institute. Picard Tools. Available: http://broadinstitute.github.io/picard/

602 50. García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, Tarazona S, Dopazo J,
603 Meyer TF, Conesa A. 2012. Qualimap: evaluating next-generation sequencing alignment
604 data. Bioinformatics 28:2678–2679.

605 51. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER,
606 Ding L, Wilson RK. 2012. VarScan 2: somatic mutation and copy number alteration
607 discovery in cancer by exome sequencing. Genome Res 22:568–576.

608 52. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017.
609 ModelFinder: fast model selection for accurate phylogenetic estimates. Nature Methods
610 14:587–589.

611 53. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and
612 Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol Biol
613 Evol 32:268–274.

614 54. Minh BQ, Nguyen MAT, von Haeseler A. 2013. Ultrafast approximation for phylogenetic
615 bootstrap. Mol Biol Evol 30:1188–1195.

616 55. raster: Geographic Data Analysis and Modeling version 3.0-12 from CRAN.

617 56. Jombart T. 2008. adegenet: a R package for the multivariate analysis of genetic markers.
618 Bioinformatics 24:1403–1405.

619 57. Treangen TJ, Ondov BD, Koren S, Phillippy AM. 2014. The Harvest suite for rapid core-
620 genome alignment and visualization of thousands of intraspecific microbial genomes.
621 Genome Biol 15:524.

622 58. Stamatakis A. 2014. RAxML Version 8: A tool for phylogenetic analysis and post-analysis
623 of large phylogenies. Bioinformatics 30:1312-1313.

624 59. R Core Team. 2018. R: A language and environment for statistical computing. R
625 Foundation for Statistical Computing, Vienna, Austria.

626 60. UNEP-WCMC. 2020. Protected Area Profile for Ngorongoro Conservation Area. World
627 Database of Protected Areas. Available: www.protectedplanet.net

628 61. Wickham H. 2009. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New
629 York.

630 62. Pebesma E. 2018. Simple features for R: Standardized support for spatial vector data. The R
631 Journal 10:439–446.

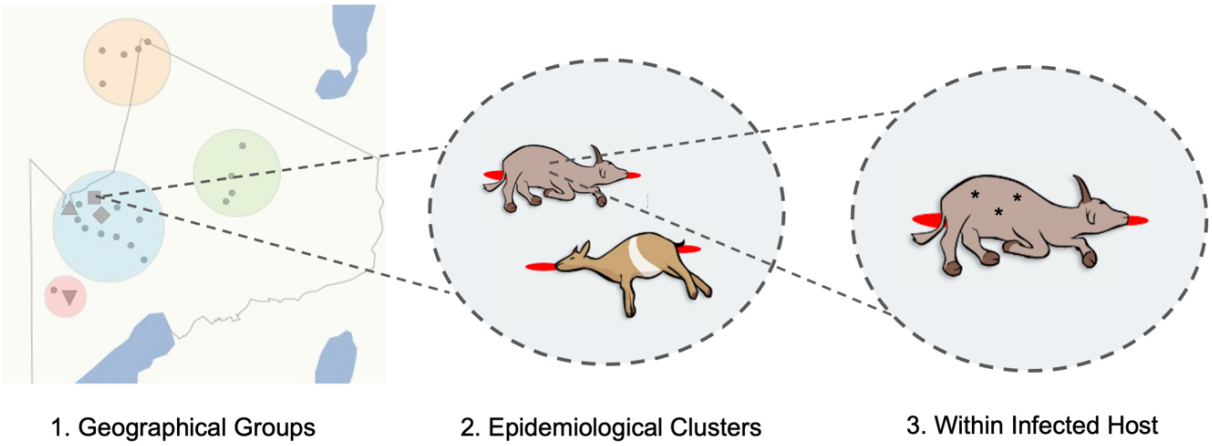
632 63. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. 2017. ggtree: an R package for visualization
633 and annotation of phylogenetic trees with their covariates and other associated data.
634 Methods Ecol Evol 8:28–36.

635 64. Ciccarelli FD, Doerks T, Mering C von, Creevey CJ, Snel B, Bork P. 2006. Toward
636 automatic reconstruction of a highly resolved Tree of Life. Science 311:1283–1287.

637 65. Hsieh TC, Ma KH, Chao A. 2016. iNEXT: an R package for rarefaction and extrapolation
638 of species diversity (Hill numbers). Methods Ecol Evol 7:1451–1456.

639

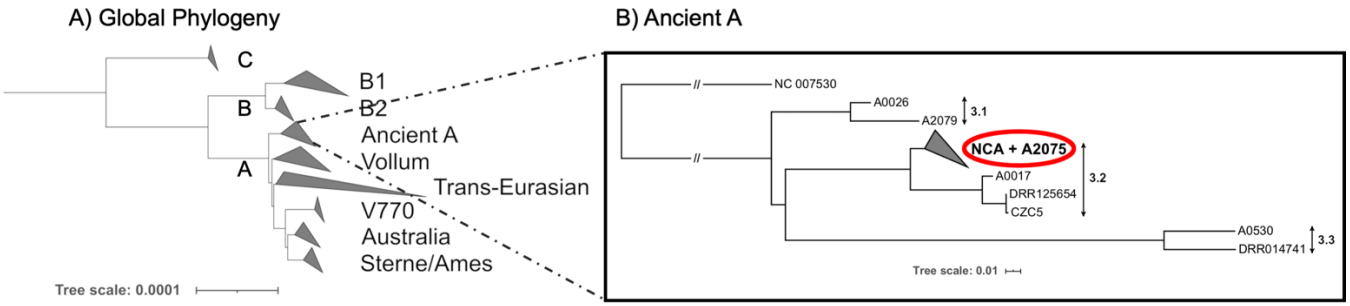
640 # Figures



642 **Fig 1. Hierarchical levels at which the genomic diversity of *Bacillus anthracis* was studied.**
643 1. Sequenced isolates originating from livestock carcasses sampled throughout the anthrax
644 hyperendemic Ngorongoro Conservation Area (NCA), northern Tanzania, (shown as grey dots in
645 panel 1) were categorized into four geographical groups (colored circles); 2. A subset of these
646 isolates were from spatio-temporally linked pairs of carcasses (n = 4, represented by grey shapes
647 in panel 1), referred to as 'epidemiological clusters'; 3. Multiple isolates (n = 2-4; represented by

648 asterixis) were sequenced from individual carcasses, either originating from multiple sample
649 types (e.g. tissue and blood), and/or multiple isolates sequenced from a single sample. The shape
650 file for the NCA was provided by Tanzania National Parks (TANAPA) (60).

651



652

653 **Fig 2. Phylogenetic position of *Bacillus anthracis* isolated from the Ngorongoro**
654 **Conservation Area within the global population.** A) Global phylogeny of *B. anthracis*,
655 showing the major clades (A, B, and C) and sub-lineages. This tree was estimated based on a
656 core single nucleotide polymorphism (SNP) phylogeny of 80 publicly available genomes (Table
657 S2). B) Maximum likelihood phylogenetic tree of the Ancient A lineage (Cluster 3 based on
658 Bruce et al., 2019). All isolates from the Ngorongoro Conservation Area (NCA) form a
659 monophyletic lineage within Cluster 3.2, along with the publicly available isolate A2075
660 (SRR2968187) isolated in 1999 from a baboon in central Tanzania. This tree was inferred using
661 the general time reversible (GTR) model of nucleotide substitution, using the Ames Ancestor
662 reference genome (NC_007530) as an outgroup. Tree scales reflect number of substitutions per
663 site.

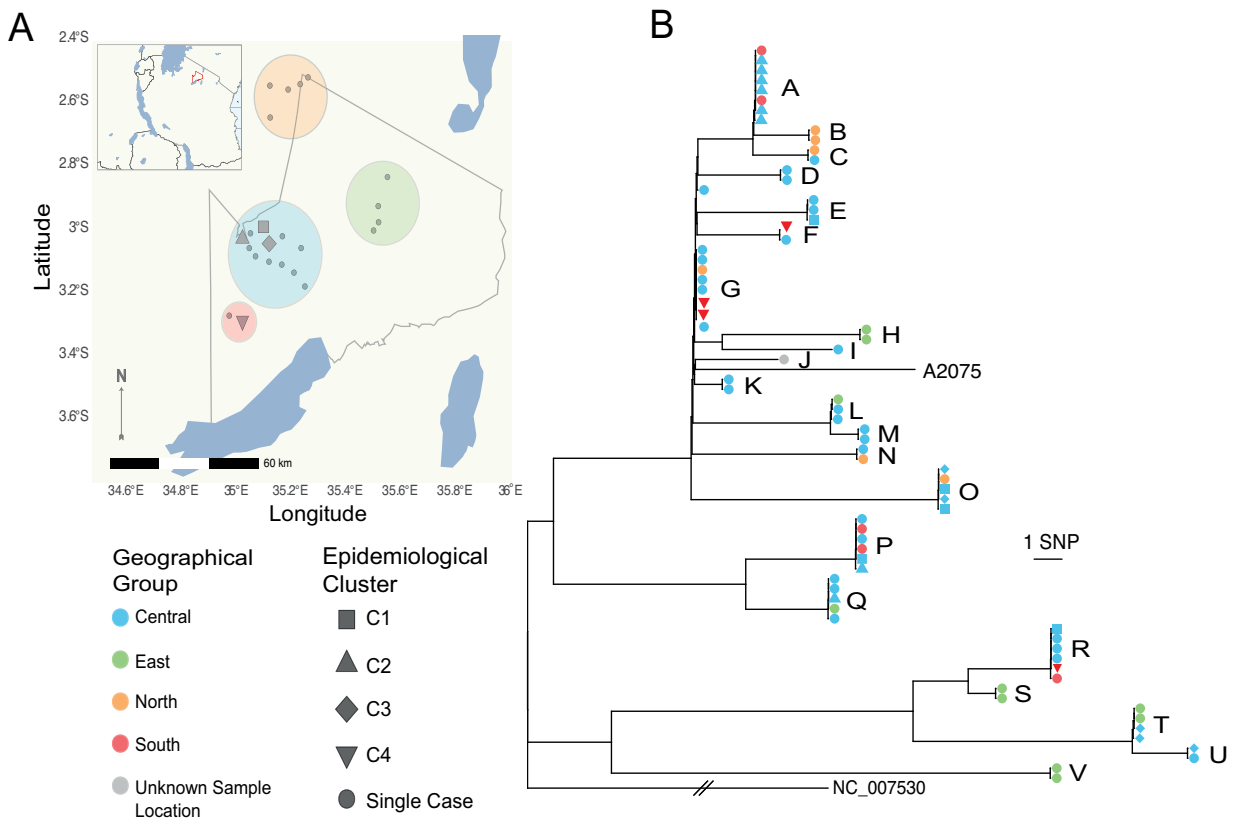
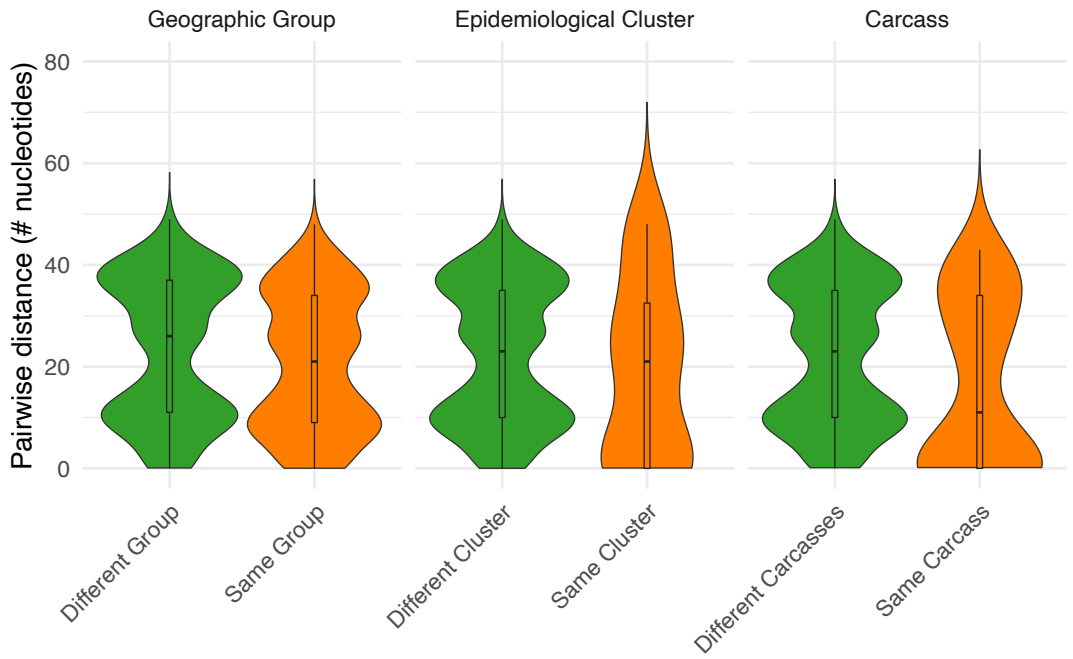


Fig 3. Phylogeography of *Bacillus anthracis* in the hyperendemic area of the Ngorongoro Conservation Area, Tanzania. A) Spatial distribution of carcasses from which *B. anthracis* isolates were obtained. The map outlines the Ngorongoro Conservation Area (NCA) and shows its location in northern Tanzania (outlined in red in inset). Carcasses were assigned to four geographical groups within the NCA based on spatial proximity, shown by colored circles. B) Maximum likelihood tree estimating the phylogenetic relationship among *B. anthracis* isolates from the NCA. This tree is based on an un-gapped alignment of 125 high quality core single nucleotide polymorphisms (SNPs) across the whole chromosome, rooted to the Ames Ancestor reference sequence (NC_005730) and including the publicly available isolate A2075 (SRR2968187) originating from central Tanzania. Using the more closely related isolates from Cluster 3.2 as an outgroup produced the same root position. Isolates are colored on the tree based on their collection site (geographical group) within the NCA. Epidemiological clusters of cases

677 (pairs of carcasses sampled from the same or neighboring households on the same or consecutive
678 days) are distinguished by symbol shape. Letters distinguish the 22 unique genotypes (SNP
679 profiles) detected. Isolate-labelled versions of these figures can be found as Fig. S4 and S5 in the
680 Supporting Information. The base earth, river and lake data for the map were downloaded from
681 Natural Earth (https://www.naturalearthdata.com/). Figure was plotted in R v.3.6.1 with *ggplot2*
682 (61), with the addition of the *sf* (62) and *ggtree* (63) packages.
683



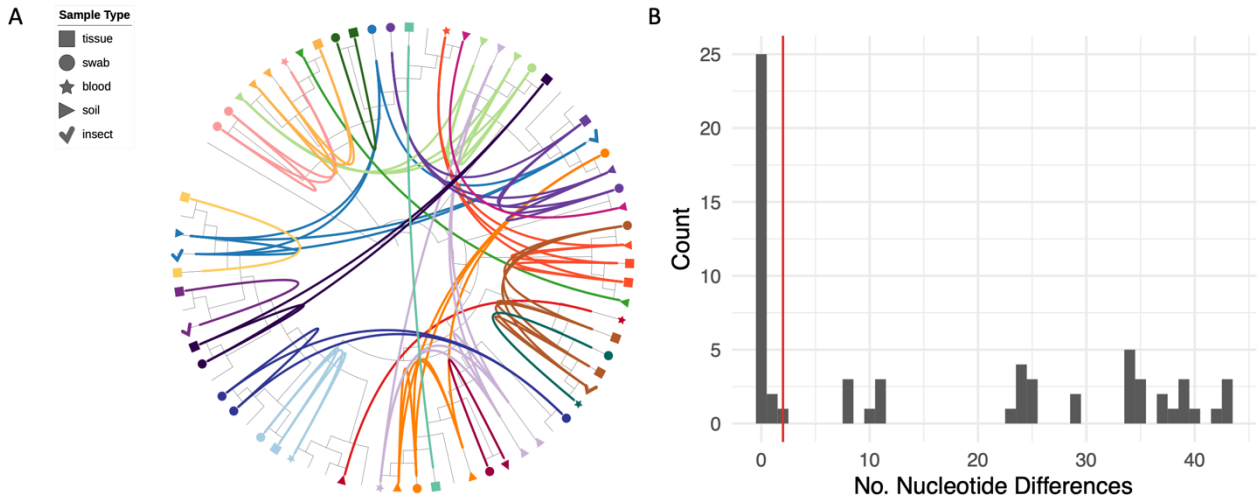
684

685 **Fig 4. Comparative numbers of single nucleotide differences among *Bacillus anthracis***
686 **isolates from hierarchical spatial scales.** Violin plots comparing pairwise nucleotide (SNP)
687 differences between all *B. anthracis* isolates from the Ngorongoro Conservation Area versus
688 SNP differences between isolates from i) within the same geographical group; ii) within the same
689 epidemiological cluster (but not from the same carcass); and iii) within a single carcass. Central

690 boxplot shows median and interquartile range, with whiskers showing minimum and maximum

691 values.

692



693

694 **Fig 5. Within-host diversity of *Bacillus anthracis* among livestock in the Ngorongoro**

695 **Conservation Area.** A) Circularized maximum likelihood tree – based on high quality core

696 single nucleotide polymorphisms (SNPs) – displayed as a cladogram (branch-lengths ignored),

697 rooted to Ames Ancestor reference genome (NC_005730). Isolates from the same carcass are

698 shown in the same color and are linked by inner connecting lines. Isolates without labels are

699 singletons (i.e. only one isolate sequenced per carcass site). Sample type is shown by the

700 different symbol shapes indicated in the legend. The figure was prepared using iTOL (64). For

701 labeled taxa, see Fig. S7. B) Histogram showing the relative frequency of pairwise SNP

702 differences among *B. anthracis* isolates collected from the same carcass. Red line shows 99%

703 upper limit of nucleotide differences observed among sampled pairs of genomes based on

704 simulation of within host evolution. Results suggest that almost all diversity observed within the

705 same infected host is the result of a heterogenous inoculum.

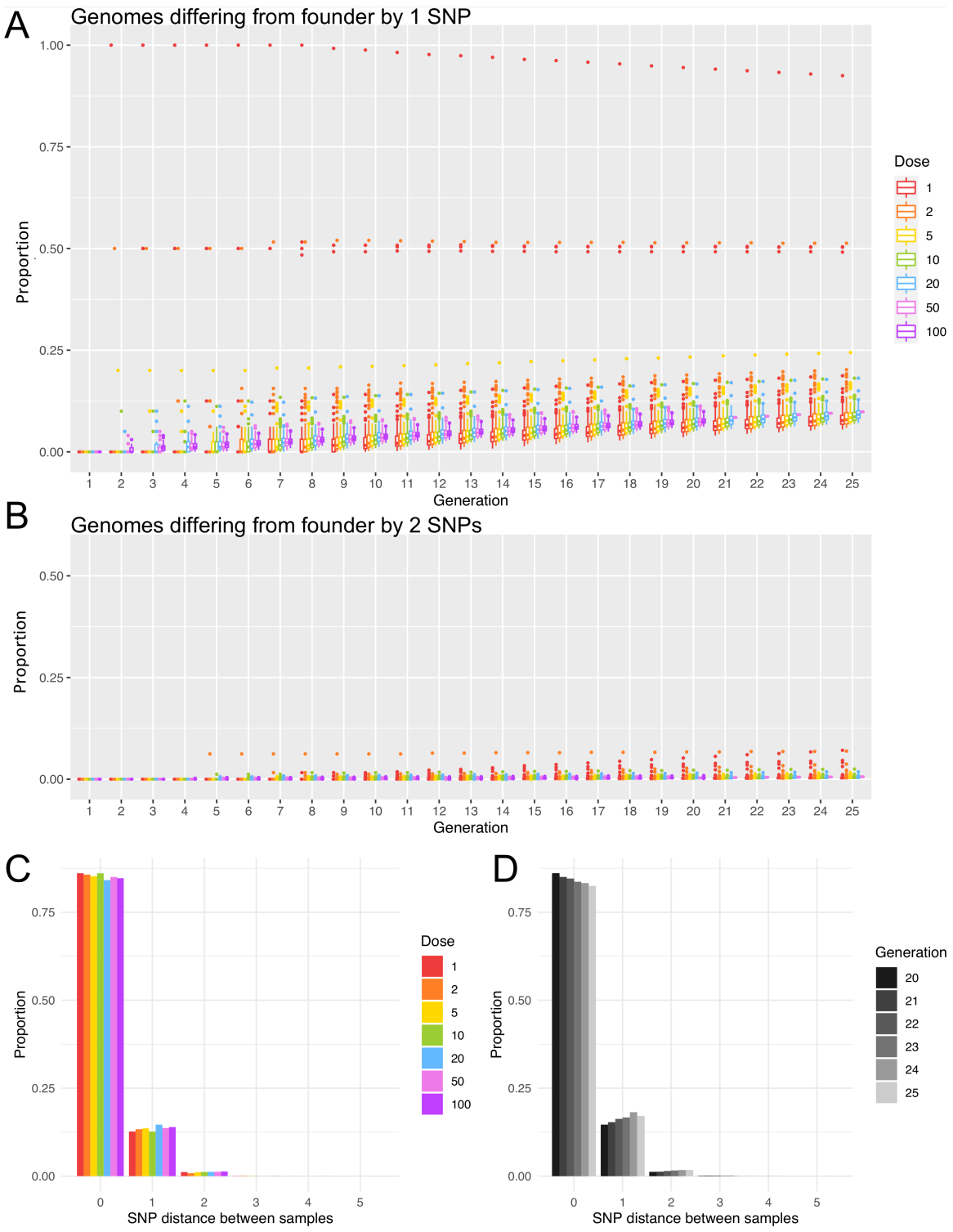
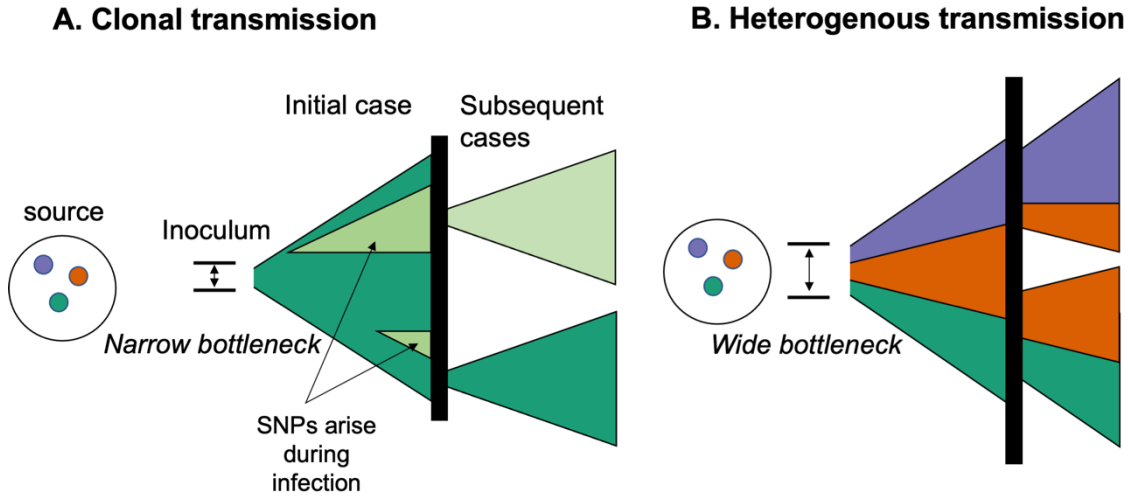


Fig 6. Simulation of within-host populations and sampling of resulting genetic diversity. Box and whiskers plots show the proportion of genomes across simulated populations that differ from the founding genome by either one nucleotide (A) or by two nucleotides (B). Boxes

710 represent the interquartile range with a line showing the median and with outliers shown as
711 points. Both boxes and outliers are colored by inoculum size (dose) according to the legend.
712 Simulations were run for 25 generations, or 20 generations for larger inoculum sizes. Greater
713 stochastic heterogeneity was observed for lower starting doses. Note that in (B), only outliers are
714 visible as the proportions observed were very low. C) Bar-plot showing the relationship between
715 inoculum size (dose) and single nucleotide polymorphism (SNP) distances between pairs of
716 sampled genomes. Bars represent average proportions of pairwise SNP differences between
717 genomes sampled from simulated within-host populations. Populations were simulated from
718 various inoculum sizes and were sampled in the 20th generation after 19 replication cycles. Fifty
719 simulations were run from an initial dose of 20, and 100 simulations were run from each of the
720 other inoculum sizes, with each simulated population sampled 100 times in each generation. D)
721 Bar-plot showing the relationship between the number of replication cycles and SNP distances
722 between pairs of sampled genomes. Populations were simulated for 25 generations and samples
723 taken in generations 20-25 are represented in shades of grey according to the legend. Proportions
724 are averaged across simulations initiated with inoculum sizes of 1, 2, 5, 10 and 20 and sampled
725 repeatedly.



726

727 **Fig 7. Conceptual framework for the acquisition of within-host diversity of *Bacillus***
728 ***anthracis*.** Colors represent different genotypes of *B. anthracis* (with multiple nucleotide
729 differences), while different shades (i.e. light green) represent single nucleotide variants arising
730 during the course of infection. Vertical black lines represent the environment (e.g. soil, water)
731 from which subsequent cases of anthrax arise through exposure to *B. anthracis* spores. A)
732 Example of infections stemming from small transmission bottleneck (inoculum of few spores),
733 resulting in primarily clonal transmission. Where the bottleneck is narrow, a limited number of
734 genotypes could comprise the inoculum, regardless of the number of variants present in the
735 environment. A small number of SNPs may arise during the course of infection (see Figs 5B and
736 6), but are unlikely to be transmitted unless they arise early during the course of infection. These
737 variants rarely differ from the founding genotype by > 1-2 SNPs (Table S5). B) Example of
738 infections stemming from a wide bottleneck (infectious dose with multiple spores), wherein
739 sufficient numbers of spores comprise the inoculum such that multiple genotypes present in the
740 environmental source may seed initial and subsequent infections. Our results strongly support
741 heterogenous transmission (B), either from single or multiple carcass sites, and a large
742 transmission bottleneck. Figure adapted from Ågren et al., 2014.

743

744 # Supporting Information

745 **Fig S1. Rarefaction curve of *Bacillus anthracis* genotypic diversity within the study area.**
746 Based on inclusion of all isolates sequenced (n = 73, left) and on all isolates that were not
747 sampled as part of an epidemiological cluster (n = 51, right), which might be expected to be non-
748 independent. Results suggest genotypes within this population have been exhaustively sampled

749 (i.e. that further sampling would not be expected to reveal additional genotypes). Figure
750 generated in R package iNEXT (65).

751 **Fig S2. Histogram showing the relative frequency of pairwise nucleotide (SNP) differences**
752 **among *B. anthracis* isolates.** Includes 73 isolates from the Ngorongoro Conservation Area,
753 northern Tanzania.

754 **Fig S3. Scatter plots showing the number of nucleotide differences as a function of**
755 **geographic difference between the sampling locations.** Geographic distance (in meters) is
756 shown on the x-axis versus nucleotide differences on the y-axis, with each point representing a
757 pair of isolates. A) All *Bacillus anthracis* isolates from the study area. B) The same relationship
758 is observed when limited to isolates from the dominant clade; this was done in order to account
759 for deeper divergences potentially obscuring patterns.

760 **Fig S4. Locations of carcasses sampled within the Ngorongoro Conservation Area,**
761 **northern Tanzania.** Each carcass is assigned a unique identifier (Table S1). Epidemiological
762 clusters are the same as those explained for Fig 3 and Fig S5.

763 **Fig S5. Phylogenetic relationship among *Bacillus anthracis* isolates from the Ngorongoro**
764 **Conservation Area.** Estimated through maximum likelihood, based on high quality core SNPs.
765 Geographic groups and epidemiological clusters are the same as those explained for Fig 3. Each
766 isolate is attributed to a carcass ID, with the sample type indicated following the underscore (B =
767 blood, D = soil, I = insect, S = swab, T = tissue). A number follows the sample type if more than
768 one isolate was sequenced from the same sample.

769 **Fig S6. Genotypes of *Bacillus anthracis* observed in isolates from within and between pairs**
770 **of carcasses from the same epidemiological clusters (C1-4).** Genotype letters correspond to

771 those in Fig 3 and Fig S5. Individual carcasses are numbered /1 or /2. In cluster C3, two isolates
772 were from a soil sample (C3/1) collected at the same household as the two cases (C3/2 and
773 C3/3); genotype T from this soil sample was shared with an isolate from C3/2. Otherwise, only
774 in C2 was there evidence of a shared genotype between pairs of carcasses (genotype A). Thus,
775 the level of sampling conducted here (1-4 isolates per carcass) did not produce evidence for the
776 same combinations of genotypes being found among linked carcasses.

777 **Fig S7. Within-host diversity of *Bacillus anthracis* isolated from livestock in the**
778 **Ngorongoro Conservation Area of northern Tanzania.** This circularized maximum likelihood
779 tree – based on high quality core single nucleotide polymorphisms – is displayed as a cladogram
780 (branch-lengths ignored). Isolates from the same carcass are shown in the same colour and linked
781 by inner connecting lines. Isolates in black are singletons (i.e. only one isolate sequenced per
782 carcass site). The figure was prepared using ITOL (64).

783 **Fig S8. Proportion of simulated within-host populations identical to the inoculating genome**
784 **over 25 generations.** Populations were simulated from homogenous inoculating doses of
785 varying size (A-G) and each line tracks a single simulated population through generations.
786 Represented are 100 simulations run for 25 generations from doses 1, 2, 5 and 10; 50 simulations
787 run for 25 generations (dose 20); 100 simulations run for 20 generations from doses 50 and 100
788 and 7 simulations run for 25 generations from inoculum dose of 50.

789 **Fig S9. Proportion of simulated within-host populations differing from the inoculating**
790 **genome by one nucleotide (SNP).** Populations were simulated from homogenous inoculating
791 doses of varying size (A-G) and each line tracks a single simulated population through
792 generations. Represented are 100 simulations run for 25 generations from doses 1, 2, 5 and 10;

793 50 simulations run for 25 generations (dose 20); 100 simulations run for 20 generations from
794 doses 50 and 100 and 7 simulations run for 25 generations from inoculum dose of 50.

795 **Fig S10. Proportion of simulated within-host populations differing from the inoculating**
796 **genome by two nucleotides (SNPs).** Populations were simulated from homogenous inoculating
797 doses of varying size (A-G) and each line tracks a single simulated population through
798 generations. Represented are 100 simulations run for 25 generations from doses 1, 2, 5 and 10;
799 50 simulations run for 25 generations (dose 20); 100 simulations run for 20 generations from
800 doses 50 and 100 and 7 simulations run for 25 generations from inoculum dose of 50.

801 **Table S1. Metadata and sequence quality metrics of *Bacillus anthracis* isolates included in**
802 **this study.** Each isolate is attributed to a carcass ID, with the sample type indicated following the
803 underscore (B = blood, D = soil, I = insect, S = swab, T = tissue). A number follows the sample
804 type if more than one isolate was sequenced from the same sample.

805 **Table S2. Publicly available *Bacillus anthracis* isolates included to contextualize those from**
806 **the Ngorongoro Conservation Area.** Sheet 1: Publicly available global collection of assembled
807 *B. anthracis* genomes. Sheet 2: Publicly available sequence data for Ancient A isolates, available
808 on SRA.

809 **Table S3. Identical isolates found from different geographical groups.**

810 **Table S4. Identical or nearly identical isolates from carcasses sampled several months**
811 **apart.**

812 **Table S5. Single nucleotide polymorphism (SNP) distances among pairs of isolates sampled**
813 **from simulated within-host populations of *Bacillus anthracis*.** Proportion of pairs of isolates

814 in evolved populations with different SNP distances across varying initial inoculum size (dose),

815 sampled in generations 20 and 25, and mean SNP differences across sampled pairs.

816 **File S1. Supplementary methods and results.**