Coalescence modeling of intra-infection *Bacillus anthracis* populations allows estimation of infection parameters in wild populations

W. Ryan Easterday[1], José Miguel Ponciano[2], Juan Pablo Gomez[3], Matthew N. Van Ert[4], Ted Hadfield[4], Karoun Bagamian[4,5], Jason K. Blackburn[4,5], Nils Chr. Stenseth[1*] and Wendy C. Turner[6*]

[1] CEES, Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, Blindernveien 31, 0317 Oslo, Norway
[2]Department of Biology, University of Florida, Gainesville, FL 32611
[3] Departamento de Química y Biología, Universidad del Norte, Barranquilla, Colombia.
[4]Emerging Pathogens Institute, 2055 Mowry Road, University of Florida, Gainesville, FL 32611
[5]Spatial Epidemiology & Ecology Research Laboratory, Department of Geography, University of Florida, Gainesville, FL 32611
[6]University at Albany, State University of New York, 1400 Washington Avenue, Albany, NY 12222

Author contributions: WRE, WCT and NCS designed the study. WRE and WCT carried out the experimental parts of the study. MVE, TH and KB did DNA based work including extraction and genotyping. JMP and JPG did coalescence modeling and detailed methods.

All authors contributed to writing the manuscript.

*To whom correspondence may be addressed. Email: n.c.stenseth@ibv.uio.no or wcturner@albany.edu

Abstract

Bacillus anthracis, the etiological agent of anthrax, is a well-established model organism. For *B. anthracis* and most other infectious diseases, knowledge regarding transmission and infection parameters in natural systems is, in large, comprised of data gathered from closely controlled laboratory experiments. Fatal, natural anthrax infections transmit the bacterium through new host-pathogen contacts at carcass sites, which can occur years after death of the previous host. For the period between contact and death, all of our knowledge is based upon experimental data from domestic livestock and laboratory animals. Here we use a non-invasive method to explore the dynamics of anthrax infections, by evaluating the terminal diversity of *B. anthracis* in anthrax carcasses. We present an application of population genetics theory, specifically coalescence modeling, to intra-infection populations of *B. anthracis* to derive estimates for the duration of the acute phase of the infection and effective population size converted to the number of spores establishing infection in wild plains zebra (*Equus quagga*). Founding populations are small, a few spores, and infections are rapid, lasting roughly between one and three days in the wild. Our results closely reflect experimental data, showing small founding populations progress acutely, killing the host within days. We believe this method is amendable to other bacterial diseases from wild, domestic and human systems.

Significance statement

This study is the first to apply coalescence modeling to a "slowly evolving" bacterial pathogen, *Bacillus anthracis,* to derive estimates of infection durations and founding population sizes from natural anthrax mortalities. Although coalescence modeling has been applied to highly mutable chronic pathogens (i.e. HIV), to date methodological hurdles prevented its wider

application. Our findings show it is possible to obtain pathological data from infections, *post-hoc*, which may be applicable to other pathogens and settings, including clinical. Given their higher resolution, microsatellites will remain useful in shorter evolutionary timeframe studies.

Introduction

Questions regarding pathology of microorganisms are often addressed using animal models. Since the validation of germ theory (using *Bacillus anthracis*) (1), animal models have been used to elucidate various parameters of infection, such as infectious dose, strain lethality, disease pathology, and host immune response (2, 3). In most studies inbred, small-animal lines are used where age, sex, diet and other variables are controlled to reduce immune response variation among individuals. Yet, it is difficult to assess to what degree these controlled studies reflect how these infectious agents behave in natural hosts. This is due to variation in immune response within heterogeneous host populations where genetic and life history variation can affect the outcome of an infection (4). Furthermore, use of natural hosts in pathological studies can be, in practice, impossible, due to necessary permissions, facilities and ethical considerations. As a result, disease pathology data are lacking in most large, wild hosts and leaves general pathological questions regarding these species, open.

The gram-positive, spore-forming bacterium *Bacillus anthracis* causes anthrax. An acute infection, anthrax can start via several routes of infection: inhalational, cutaneous, ingestional and injection. The pathogen occurs globally where its main hosts are large ungulates, yet most mammals and even birds can be susceptible (5, 6). *Bacillus anthracis* is an "obligately lethal pathogen," where the host must die for transmission to occur. In some anthrax endemic areas, transmission may be enhanced with the involvement of biting flies and blowflies (7). Yet, regardless of these other types of transmission, anthrax associated with grazing at carcass sites by new hosts is the backbone of its epidemiology across systems (5, 8).

According to Glomski et al., ingestional anthrax infections in mice can start in the upper gastrointestinal tract, associated with previous damage to the epithelium, or in the lower gastrointestinal tract, within the lymphatic tissue of the oropharynx or Peyer's patches, respectively (9). Stimulation of phagocytic cells, such as dendritic cells and macrophages to engulf spores via the classic complement pathway (CCP) plays an important role in establishing the infection. Interaction between BlcA glycoprotein, a major structural component of the *B. anthracis* exosporium (10), and complement component C1q stimulate both entry into epithelial cells and further activation of CCP, beginning the complement system cascade, marking them for uptake by phagocytic cells providing carriage across the epithelium to adjacent lymphatic tissues (11). After passage past the epithelium, the disease seems to progress very similarly, regardless of the initial route of infection. Spores germinate to vegetative cells, which proliferate and spread through the draining lymphatic system, notoriously involving the spleen, and shortly thereafter becomes a systemic infection. Hemorrhaging from orifices occurs around the time of death releasing *B. anthracis* into the soil and inducing sporulation allowing the pathogen to survive for years in the environment (8).

In Etosha National Park, Namibia, anthrax has been monitored, but not managed for roughly 40 years; throughout which, an effort was made to sample all discovered mortalities. Plains zebra (*Equus quagga*) are the most common host for *B. anthracis* in Etosha. Most of these infections likely occur after ingesting spores while grazing at anthrax carcasses sites (8), and not from drinking contaminated water (8) nor from inhalation of spores (12). Anthrax mortalities in zebra peak during the rainy season, where enhanced production of forage occurs at nutrient rich carcass sites (13). Although the majority of the zebra in Etosha have trace levels of antibodies against *B. anthracis* (indicating a high exposure rate) (14), disease incidence remains quite low,

even in outbreak years (< 5%: per communication with Steve Bellan), implying few actually succumb to the infection (15).

Our previous study described increased exposure to high concentrations of the pathogen increases the probability of infection (8). Experimentally, high doses are used to induce gastrointestinal lethal infection in various ungulates, 10s-100s of millions of spores (8). Which is in contrast to the injection route, where LD_{50s} are only tens to hundreds of spores (5), showing low doses in certain instances can lead to fatal infection.

To investigate these dynamics in nature, we isolated 30 individual colony forming units (CFUs) from 11 naturally occurring zebra mortalities and genotyped the 330 isolates using Multi-Locus VNTR Analysis (MLVA) and single nucleotide repeat (SNR) data, as these markers mutate quickly enough to allow within-host resolution. In conjunction, we conducted a mutation rate experiment to calculate the average number of mutations/gene/generation (μ), treating each VNTR or SNR as a gene. We then designed a joint Maximum Likelihood (ML) approach for the Coalescent process (16) under constant and variable effective population size (17) leveraging the experimental data and the carcass genotyping data to estimate the Time to the Most Recent Common Ancestor (TMRCA) and effective population sizes (N_e) starting a given infection. The full mathematical and statistical approach detailed in the appendix uses recent theory (18, 19), algorithms (17) and ML techniques using Markov chain Monte Carlo (MCMC) for hierarchical models (20-22).

Results

Genotype data

SNR and MLVA data yielded 43 unique genotypes from 11 carcasses (30 isolates per carcass) (SI Appendix Figure 3). All data is available per request from either corresponding author, Stenseth or Turner.

Laboratory Experiment

Assuming a constant population size across the laboratory experiment, the ML estimate for the average number of mutations separating a sample size of two genes, $\hat{\theta}$= 0.46 (CI: 0.09,1.42), using the DC methodology described in Methods. Noting that the average number of mutations that separates two genes, θ, is defined in terms of the mutation rate μ and the effective population size N_e as $\theta = 2N_e\mu$, we then used MCMC to sample from the conditional distribution of the TMRCA given the ML estimate of $\hat{\theta}$, stored the median TMRCA and computed N_e = 126.5 (CI: 101.3,181.3) by dividing the duration of the experiment in generations (N=214) by that median. The mutation rate per gene, per generation, was then computed as $\mu = \hat{\theta}/2N_e$ = 0.002 (CI: 0.0005,0.004).

Carcass Sampling

Assuming constant population size from the zebra carcasses, estimates of θ varied between 0.28 and 1.1 and thus assuming a mutation rate μ = 0.002, the effective population size of *B. anthracis* varied between 68.3 and 266.7. Time to most recent common ancestor varied between 21.4 and 80.7 days (Figure 1).

The exponential model gave radically different results. In that model, it is assumed that the effective population grows exponentially from past to present at a rate β. Under the Coalescent process, this exponential growth model for the effective population size is formulated as a change from the present (zebra's time of death) to the past till the time of infection by a "founder" *B. anthracis* population using $N_e(t)=N_e(0)e^{-\beta t}$. In this model, θ changes over time according to $\theta(t) = 2N_e(t)\mu$ (see full description of the model in the Methods). Estimates of the *B. anthracis* population θ at the moment of zebra death is given by the value of θ at time 0 and is denoted θ_0 . Its estimates for each zebra varied between 1.88 and 2.42, with β values ranging between 0.35 and 1.61 (see Table 1). The effective population size of the founder *B. anthracis* population (*i.e.*, at the beginning of the infection) is denoted as $N_e(1)$ (see Methods) and was estimated to range between 193.25 and 295.38 (see Table 1). *Ne* values are converted to colony forming units (cfu) using the N_e scaling given by the mutation rate experiment. Since this experiment was started with 1 cfu, we then scaled effective population sizes assuming that N_e of 126.5 = 1 cfu. The cfu's at the moment of death estimated for all sampled zebras ranged between approximately 1 and 6 (Table 1). The estimated TMRCA from the Coalescent model was used as an estimate of the elapsed time from the moment of infection with a founder *B. anthracis* population until death (see Methods). This estimate varied between 0.73 and 2.61 days for all zebras (Figure 1). Full results of the estimates of θ, β, N_e and TMRCA and confidence intervals for each parameter are shown in Table 1. Finally, model selection through Likelihood ratio tests showed the exponential population growth model was a better fit to the data for all zebras (p-value < 0.0001 in every case).

Discussion

Our best results, not surprisingly, were from the exponential model, as this most closely resembles the population growth dynamics of *B. anthracis*. From these data, we show estimates of parameters of lethal anthrax infections in free-ranging wildlife *post-mortem*. Experimentally, infections have a short duration of infection and via injection models, low infectious doses (23). Somewhat similar studies have estimated duration of infection for chronic and highly mutable viral pathogens, namely HIV (24). Yet, we believe we are the first to use this method to estimate both duration of infection and infecting founding population size on a slow-evolving, acute, bacterial pathogen (25). It should be noted the model used here applies to *Bacillus anthracis* as the assumptions we make reflect the biology of this highly-clonal pathogen. Stratilo et al. were the first to describe the use of SNRs to characterize diversity within infections (26). To date it is likely the only developed typing system using SNRs (27).

Population dynamics

Bacillus anthracis populations fluctuate through transmission and infection stages. During an infection the population increases exponentially and afterwards goes through three transmission bottlenecks (Figure 2 below) to start an infection in a new host. These bottlenecks occur in succession, the first is a slow process of spore decay at carcass sites. This decay may be augmented slightly by some vegetative activity during this telluric process (28), nevertheless the overall trend is decay (fig 2C-D), a process taking years (8). The other two bottlenecks occur

during the infection process, first upon ingestion of a subset of spores (ingested dose) from a carcass site and finally, a bottleneck as a portion of the ingested population that establishes the infection (founders), which we calculate here in this study (Table 1).

Grazing and exposure to *B. anthracis* (Fig 2, A)

While many vertebrates are suitable hosts for *B. anthracis*, the foraging behavior and overall ecology of many herbivores lends them to be the major hosts and maintainers of anthrax in natural settings. Here ingestional anthrax, contracted from grazing at contaminated carcass sites (13), is purportedly the most common pathway of infection in wild and domestic ungulates, although other routes of transmission may occur (7, 8, 29, 30). For *E. quagga* in Etosha, grazing and ingestion of spores via contaminated plants and soil represents the largest hazard. It is difficult to know how strong of a bottleneck occurs between the ingested dose and the infecting dose, as the dose ingested is likely to be highly variable depending on site age and host behavior. However, simulation models of zebra foraging behavior indicate there is a high probability of ingesting doses up to 10^6 spores with even a bite or two of grass at a carcass site within the first two years(8). Over 5 years of simulations, there remained a spike in the probability of ingesting doses up to 10^5-10^6; doses higher than this were highly improbable.

Establishment of the infection (Fig. 2, B)

After ingestion, the process of infection establishment begins. For mouse gastrointestinal animal models, two major locations, the oropharynx (when epithelium is damaged) and/or Peyer's Patches, are tissues commonly associated with *B. anthracis* entry from the lumen into the

body (9). In wild ungulates, infection establishment has been speculated to be enhanced through damaged tissues caused by rough forage (31, 32) or gut parasites such as helminths due to higher activity of immune cells at these wound sites (32). Entry occurs through phagocytosis of spores by macrophages, carriage across the epithelium and transport to lymphatic tissue. After phagocytosis, spores germinate and the vegetative cells escape the phagosome starting the infection (33). High proportions of spores can germinate within hours, but can also be quite staggered depending on germinates present (34).

Although anthrax establishes via several routes of infection, crossing the epithelium is typically mediated through macrophages, and from our data and in accordance with Lowe et al. (35), *B. anthracis* incurs a large population bottleneck starting the infection. Parsimoniously, our data suggest a small population can result in these animals and progress quickly to a lethal infection. The majority of the subsequent population diversity seems to be arising in-host, hence there are very similar diversity patterns among infections. Likewise, Lowe et al. suggests a similar mechanism creating a bottleneck for an intranasal anthrax model, where a substantial population bottleneck occurs between the inoculum and the founding population in the nasal mucosa-associated lymphoid tissue (NALT)(35).

For anthrax, route of infection greatly affects the necessary dose to reach an LD_{50}. This is especially true between oral and injectional routes, where the epithelium acts as an effective barrier to infection. For instance, de Vos reports that kudu (*Tragelaphus strepsiceros*) ingestional lethal doses were estimated at 1.5×10^7 (range 1×10^6 to 6.5×10^7) while a parenteral (injected) dose of 250 cells proved fatal to impala (*Aepyceros melampus*) (36). These data also reflect trends for sheep where lethality for ingestional anthrax requires much larger doses and only tens of cells required via injection (23). By our estimates, the founding population reflects the number

of spores which crossed the epithelium and successfully germinated to start the infection. Despite our estimated low number of spores, large doses of ingested spores may be required to start gastrointestinal anthrax infections. Where blcA on the outmost coat stimulates the classical complement system (11) a high dose might be needed to produce an adequate innate immune response to stimulate macrophages and dendritic cells to take up spores marked with C3 fragments. Strikingly, infectious doses among zebras in this study were very similar which reflects supathogen diversity and suggests some common pathology for *B. anthracis* and/or a shared trait among the individual zebra mortalities, such as genetic, behavioral, or life history, including previous exposure.

The success of using coalescence modeling to estimate N_e and TMRCA depends on having enough genetic resolution within the sampled population. This means having sampled enough individuals from a given population in combination with a high enough diversity, which corresponds to mutation rate. Although pathogens such as *B. anthracis, Yersinia pestis* and others are often referred to as "highly clonal" or "slowly evolving" it is important to make some distinctions. These pathogens are often classified this way due to high sequence similarity in coding regions, yet mutations such as indels (including VNTRS/SNRs), and genomic rearrangements are ignored in this classification. This is especially true with the use of genome sequencing for population studies, where most often resequencing and aligning to a reference are used, which often technically have hurdles in assembling larger VNTRs and ignore rearrangements in favor of reference synteny. Yet, longer read technology and de novo alignment will make these data available. In conclusion, this method may be quite amendable to other disease systems and even clinical settings, given these types of markers (VNTRs and SNRs) are used and may yield valuable information for curtailing disease transmission.

Methods

Study area

This study was conducted using isolates of *B. anthracis* collected from anthrax carcasses in central Etosha National Park, Namibia from 2008-2012. Anthrax is endemic in Namibia, and Etosha National Park has regular annual outbreaks of anthrax recorded primarily in grazing herbivores (37, 38). More than 50% of anthrax cases recorded are of plains zebras (*Equus quagga*), and among the herbivorous species, zebras show the strongest propensity for foraging on grasses at anthrax carcass sites (13).

Isolation of B. anthracis from blood swabs

Culture and isolation of *B. anthracis* was done at the Etosha Ecological Institute's pathogen laboratory. Dried, refrigerated carcass swabs from 11 zebra anthrax mortalities with three zebra from 2008, four from 2009, two from 2010 and two from 2012 (SI Appendix, Figure 2) were used to collect isolates for this study. Swabs were rehydrated in 1.5ml sterile distilled water and agitated occasionally for several minutes to suspend spores. Dilutions of 10^{-2}, 10^{-4} and 10^{-6} were prepared and plated on PLET agar using 5µl of each dilution and the undiluted with an additional 50µl of sterile, distilled water to spread the sample evenly over the agar. 30 isolated colonies were selected from among the plates for each carcass. If a particular morphology was in doubt as to whether or not it was *B. anthracis*, standard confirmation tests (penicillin and ϒ-phage) on a representative from that morphology were done before picking samples. Entire colonies were transferred from the culture plates to 0.5ml cryotubes containing 0.25ml PLET agar using sterile toothpicks and incubated for several days at 37C before shipping at ambient temperature to University of Florida in Gainesville.

Mutation rate experiment methods

An isolate was obtained from a blood swab from zebra carcass containing the most common genotype in Etosha (genotype 6) according to Beyer et al. (39). This isolate is from A.Br.003 (A.Br.Aust94) using Van Ert et al.'s global classification (40), and group 5.4 using a new population genomic classification (41). The zebra carcass was found on 22 February 2010 (carcass ID: EB100224-01WT). The colony was placed into 25mL of Difco nutrient broth in a 50mL tube and mixed gently in an incubator at 37°C (range 35-41°C) for 24 hours. The remaining part of the colony was transferred to a cryotube to preserve as the initial diversity for the experiment. After 24 hours, the *B. anthracis* culture in nutrient broth was diluted to 10^{-6} in sterile water. We then inoculated 1μL of 10^{-6} dilution into 60 50mL tubes each with 25mL of nutrient broth. These 60 samples were gently mixed in the incubator at 37°C for 24 hours. From these original 60 tubes, five additional serial transfers were done. 61 isolates from 60 lineages and the 1 progenitor were shipped to U of F. The starting isolate used for this experiment was sequenced and is available on GenBank (SubmissionID: SUB6568587 or BioProject ID: PRJNA590262).

DNA extraction. At the University of Florida, isolates were grown on 5% sheep blood agar for 24 – 48 hours and DNA was isolated using a modification of the method presented by Van Ert (40).

MLVA-25 genotyping. MLVA-25 genotyping was performed as described by Lista et al. (42), with minor changes in PCR chemistry and volumes to reduce genotyping costs and adaptations in primer labeling to accommodate analyses on the Applied Biosystems (ABI; Applied

Biosystems, Foster City, California) instruments. Briefly, cold start, multiplex PCR was performed using 5.0 µL reactions containing: 0.5 U/rxn Taq DNA Polymerase (Syd labs, Natick, MA), 1X Syd Taq Buffer (contains $MgCl_2$), 1X concentration of multiplex primer mix, 0.25 mM each dNTPs (Applied Biosystems, Foster City, California), and 0.5 µL of template DNA. Thermal cycling conditions were as per Lista et al. with exception of omitting the initial denaturation step (cold start polymerase). PCR products were diluted 1:40 by the direct addition of 195 µL of molecular grade water to the PCR plates and 1.0 µL of diluted product was added to 19.0 µL of a formamide/LIZ 1200 (ABI) size standard mixture (0.285 uL size standard per well) and denatured. Electrophoresis was conducted on an ABI 3730 sequencer and fragment sizes determined using GeneMapper™ software (Applied Biosystems).

SNR-4 genotyping. The four SNR loci described in Kenefic et al. (27) were amplified in multiplex The 10.0 µl PCR reactions were carried out with final concentrations of the following: 1.0 µL template DNA per reaction, 1X PCR buffer, 0.5 U per reaction *pfu* Polymerase (Agilent technologies, Wilmington DE), 3 mM $MgCl_2^*$, and 0.25 mM of each dNTP. The final primer concentrations in the reaction were 0.1 µM of HM-1, 0.15 µM of HM-2, 0.1 µM of HM-6 and 0.25 µM of HM-13. The PCR products were diluted 1:20 and 1.0 uL was mixed with 19.0 uL of a formamide/LIZ 500 (Applied Biosystems) size standard mixture (0.285 uL standard per rxn) and denatured. Fragment sizing for SNR-4 was performed on an ABI 3730 (Applied Biosystems) and array sizes were determined using GeneMapper™ software (Applied Biosystems).

MODELING APPROACH: AN OVERVIEW

In what follows we briefly overview our modeling approach using the Coalescent Process (16), our analyses rationale as well as the questions we sought to answer with them. Then, we give a detailed statistical account of our methodologies.

Here we used statistical inference for the Coalescent Process (16) to leverage the results from the serial passage culturing of *B. anthracis*, and the MLVA and SNR types sampled from the eleven zebra carcasses. In a landmark paper, Tavaré et al (43) showed how to use computational sampling methods to estimate the Time to the Most Recent Ancestor (TMRCA) from a sample of size *n* genes and the count of "segregating sites", or the number of variable loci in these genes. Critical for their inferential approach is the adoption of a mutation model. As these authors mention, a wide variety of models for the mutation process can be incorporated into the Coalescent. When the data are DNA sequences, the infinitely-many-sites model (44) may be appropriate. This model is commonly applied to sequence data (e.g. cytochrome b mtDNA used in 45 to infer ancestry) and variation at loci among the sampled genes. In this case, we refer to a gene as a sequence from an individual (or sample in our case). Specifically, these datasets consist of the sequence of nucleotides at a specific region of the genome for which individuals are variable at specific loci within the region. The number of these variable loci is the number of segregating sites, which is critical for our calculations. Furthermore, identical sequences within a group of individuals are labeled as haplotypes and their frequencies in the sample are recorded (see Fig 1 in 45).

A careful reading of Watterson (44), Ward et al. (45) and Tavaré et al (43) suggests the infinitely-many-sites model seems to be equally applicable to MLVA and SNR data structure and nature of polymorphic microsatellites. With respect to the data structure, the analogy is as follows: in our case, the equivalent to one DNA sequence haplotype is a series of the

MLVA/SNR alleles at every MLVA/SNR locus found in one sample (e.g. Appendix 1A). In what follows, we call each different sequence of MLVA/SNR alleles a MLVA/SNR haplotype. Also, just as with the mtDNA data, we also have the observed frequencies of each one of the MLVA/SNR haplotypes within the samples in each zebra. The annotated table of MLVA/SNR haplotypes and their frequencies is shown in SI Appendix table 1. In that table, n_i refers to the total number of samples for zebra *i,* (i = 1, 2… 11). For more details about the data structure and notation, see the example in the *Statistical Analyses* section below.

With respect to the biological justification of the applicability of the infinitely-many-sites model to the MLVA/SNR data set, the analogy with Watterson's setting is as follows. Watterson first assumed as his data unit, a portion of DNA specifying a single polypeptide chain of an enzyme (a functional "gene"). Recombination due to crossing over could be ignored so new alleles only result from mutation. Furthermore, the model does not require accommodating linkage and/or independence among loci. The model name, "infinitely-many-sites" assumes no two mutations ever occur at the same site (locus) so at each site, there are only two possible nucleotides: the original wild type and the mutant type. In our case, then, adopting this model assumes the inter-allelic mutations at each MLVA/SNR locus are symmetrical and identical. Although we recognize this assumption is a simple approximation of reality, it allows a clever MCMC (Markov Chain Monte Carlo) solution by Ewens and Joyce (17) (described in our *Statistical Analysis* section below) to bypass the integration over all genealogies and target the estimation of the TMRCA, while ignoring the estimation of the topology of the genealogical tree among the MLVA /SNR genes. Having a quick access to the estimation of the TMRCA allowed us to: first, estimate the TMRCA from the serial transfer experiments, calibrate this Coalescent time with real time units (in days) and estimate a laboratory effective population size and

mutation rate. Second, it allowed us to estimate the time (in days) from initial host infection to host death as the TMRCA between all the MLVA/SNR variants sampled within a single host, for each host. Third, it allowed us to carry a test of the hypothesis of within host exponential growth of the effective population size vs. the usual Coalescent assumption of constant effective population size. Infection by *B. anthracis* undergoes at least two bottlenecks driven by host resistance in specific organs (35), suggesting a model with exponential growth posterior to initial infection might be a more realistic scenario than the constant population size model. Fourth, adopting the infinitely-many-sites model allowed estimates of the effective population size of the MLVA/SNR genes upon death for each zebra. Finally, our methodology also allowed us to estimate the effective population size for these genes at the onset of host infection. In that sense, the joint estimation of the effective population size and the hypothesis test mentioned above allowed us to distinguish between two hypotheses 1) each host was initially invaded with a large *B. anthracis* load which did not grow significantly 2) zebra were initially infected with a small *B. anthracis* load, which grew fast and exponentially during infection. The comparison of the effective population sizes with the laboratory effective population size which underwent various bottlenecks, allowed us to discuss the within host population processes from the time of infection until host death.

In what follows, we delve into the mathematical modeling details, starting with the description of the model parameters and likelihood functions under both models, and detailing the Coalescent time scaling transformation to real time units.

STATISTICAL ANALYSES

Data structure and general model setting

Before setting our statistical notation, recall that here, our functional "gene" unit is the *B. anthracis* genome*,* genotyped for 25 MLVA and four SNR sites for any one sample within a zebra. For zebra 2, for example, for which there were 26 samples (our "genes"), four MLVA/SNR sites were variable (see SI Appendix 1A and B for the table presenting the raw data). These samples have 7 distinct MLVA/SNR haplotypes. Heretofore, we will simply say for zebra 2 we have 26 sampled genes and 7 MLVA/SNR haplotypes, each one with frequencies shown in Appendix 1C.

The key parameter in the Coalescent process with neutral mutations is θ*,* the average number of mutations separating a sample of size $n = 2$ genes. Furthermore, $\theta = 2N_e\mu$ where N_e is the "effective population size" and μ is the mutation rate (per gene, per generation). "N-Coalescent" time is measured retrospectively, with 0 being at present and increasing from present to past. Formally, this stochastic process is a pure death process (16), where the quantity that is "dying" is the number of distinct gene lineages, from present to past. This effective population size N_e is assumed constant over time and is defined as the size of the "population" of genes from which the samples in the present time are taken. This quantity is equal to the census population size in an idealized Wright-Fisher model (19). Although N_e is an abstract parameter, for a real biological population it is proportional to the rate at which genetic diversity is lost or gained. In the absence of natural selection and if the variation in the number of descendant genes per gene as well as the generation time are known, a census population size can be approximated (46). To date, statisticians working in this field (e.g. 19) adopt a more cautious interpretation of the effective population size and simply see it as a measure of relative genetic diversity (47, 48). In any case, this parameter (N_e) is useful, because under the Coalescent, time is re-scaled so one unit of continuous Coalescent time is equivalent to N_e generations ($2N_e$ is used in diploid

models). With that scaling, we can transform our estimated TMRCA expressed in Coalescent time units into real time units.

Several Coalescent-based methods for estimation of N_e were derived using stringent and flexible assumptions, such as constant population size, exponentially growing population size, logistic and piecewise linear. To remove the inflexible conditions imposed by adopting any time-dependent model, Palacios and Minin (19) go so far as to propose a non-parametric, stochastically varying Markov Random Field model for N_e (19). Even this last complex model formulation can be tied to a specific mathematical model of population dynamics: a translated Stochastic Gompertz diffusion model of population size growing under environmental variability (Ponciano 2018). Because most implementations of the Coalescent under variable population size can be tied to a population dynamics rationale, we opted for testing the applicability of the constant vs. the exponentially growing $N_e$as way to compromise between biological realism and estimability of parameters in the light of the data. Although most of these methodologies have been implemented and readily available software exists (e.g. "BEAST") to analyze the data under different models, these programs rely on a set of hard-coded genetic mutation models to carry the likelihood calculation by integrating the genealogy likelihood over all possible genealogies (49). Because we are mainly interested on the estimation of the TMRCA and not on the topology of the within-host genealogies, we used the approach proposed by Ewens and Joyce (17) to deal with this case to swiftly bypass the topology estimation problem. Although in their lecture notes, Ewens and Joyce only outline this approach, here we coded it de novo and extended it for the joint estimation of θ and the TMRCA (scaled to real time units) under a constant effective population size model and an exponentially growing effective population size model. The code was originally written by one of us (JMP) during a mathematical population

genetics workshop taught in 2009 by Joyce, Ewens, Krone and Ponciano at the Center for Research in Mathematics in Guanajuato, Mexico.

The joint distribution of Coalescent times

The Coalescent process is a continuous time, discrete state Markov death process, which is initiated at the present time by gathering a random sample of n genes from a population of N_e genes. Then, the process models how the number of distinct gene lineages sampled in the present decreases one at a time when we traverse time from the present to the past. When two genes sampled today find a common ancestor j generations back into the past, we say a "coalescence" has occurred. These "Coalescent events" happen until all genes in a sample have found a common ancestor. Kingman (16) and multiple authors subsequently described the mathematical properties of the retrospective and random time period elapsed since the moment one finds n genes in a sample until all of these genes have found their most common recent ancestor (TMRCA). Regardless of the assumptions about the size of N_e , TMRCA adopts a probability distribution that can be thought as the sum of all the inter-Coalescent times in a genealogy, which are all the time periods between two consecutive coalescences in a genealogy. Using stochastic processes terminology, these inter-Coalescent times are the inter event times of the Markov death process.

One attractive feature of the Coalescent model is its mathematical simplicity, which allows an intuitive understanding of the model properties and of the inter-Coalescent events using simple biological and probabilistic rationales. The number of discrete generations from the present to the past until the first coalescence occurs is modeled using a Geometric random variable where the "success" probability p is the probability that in a sample of n genes, 2 individuals find a common ancestor one generation in the past. Its complement, $1- p$ is the

probability that no coalescence occurs. Thinking of generations as independent trials, the probability of any two genes among these *n* genes finding a common ancestor *j* generations back in the past is simply $(1-p)^{j-1}p$ and the probability of their first common ancestor appearing more than r generations ago is $(1-p)^r$. The analytical expression for *p* is found as follows: The probability any two genes picked at random today have 2 different ancestors one generation back in the past is $\frac{N_e}{N_e}\left(\frac{N_e-1}{N_e}\right)=\left(1-\frac{1}{N_e}\right)$, since the first gene has N_e choices for its ancestor and the second *N-1* choices. The probability that these two genes have a common ancestor one generation back in the past (i.e., that a coalescence occurs) is then simply $1-\left(1-\frac{1}{N_e}\right)=\frac{1}{N_e}$. This fraction only gives us the value of *p* for a sample of size 2 genes. Also, note the expected number of generations until two individuals find their common ancestor is $1\ /\ (1\ /\ N_e)=N_e$. Iterating the above argument to include 3 or more genes, it is easy to see that the probability 1-*p* for a sample of *n* genes all find different ancestors one generation back in the past is

$$\prod_{i=1}^{n-1}\left(1-\frac{i}{N_e}\right)\approx 1-n_{\frac{(n-1)}{2N_e}}$$

and hence the probability of at least one coalescence occurs one generation back in the pasts is $1^{-}\left(1-\frac{n(n-1)}{2N_e}\right)=\binom{n}{2}\frac{1}{N_e}$. Denoting the inter-Coalescent, geometrically distributed, random time between k and k - 1 gene ancestors as U_k, it follows that

$$P(U_k>r)=\left[\prod_{i=1}^{k-1}\left(1-\frac{i}{N_e}\right)\right]^r\approx\left(1-\frac{k(k-1)}{2N_e}\right)^r$$

for constant population size. Now, if N_e is large relative to $(n-1)/2$, Coalescent events will occur rarely: many generations would elapse before a coalescence occurs. It then makes sense to re-scale time using a continuous scale instead of discrete generations by measuring it in units of N_e so that $r = N_e t$ Coalescent time units (*e.g.* one Coalescent time unit is equivalent to N_e generations). Applying this re-scaling is achieved by computing the limit

$$\lim_{N_e \to \infty} P(U_k > t) = \lim_{N_e \to \infty} \left(1 - \frac{k(k-1)t}{2N_e t}\right)^r = e^{-\frac{k(k-1)}{2}t} \quad .$$

Thus, measured in continuous time, the inter-Coalescent time between *k* and *k - 1* gene ancestors can be modeled using an exponential distribution with rate $\binom{k}{2}$. The TMRCA can be simply modeled as a sum of exponentially distributed inter-Coalescent times. Using the Markov property, the joint probability distribution of the inter-Coalescent times is simply written as the product of all the inter-Coalescent exponential distributions.

To set notation as well as visualize these inter-Coalescent times, we plotted a realization of a genealogy under the Coalescent process assuming at present, a sample of n=7 genes was gathered (SI Appendix Figure 1). In that graph, the u_i's denote realizations of the (random) inter-Coalescent times and t_i's denote the accumulated time, from the present to the past. Accordingly, $u_k = t_{k-1} - t_k$ or equivalently, $t_{k-1} = u_k + t_k$ under a model of changing effective population size, denoted $N_e(t)$, the probability density function (pdf) of the inter-Coalescent times is no longer exponential. Instead, the pdf of each inter-Coalescent time is

$$\Pr(u_k | t_k) = \frac{k(k-1)}{2N_e(u_k + t_k)} \times \exp\left\{-\int_{t_k}^{u_k + t_k} \frac{k(k-1)}{2N_e(t)} dt\right\},$$

(50) and their joint pdf is simply written as the product of these densities, for $k = n, n-1, \ldots, 2$. When it is assumed the population grows exponentially from past to present at a rate β (or

alternatively, decays exponentially from present to the past), expressed as $N_e(t) = N_e(0)\mathrm{e}^{-\beta t}$ then

$$\mathrm{p_r}(u_k|t_k) = \frac{k(k-1)}{2N_e(u_k+t_k)} \times \exp\left\{-\frac{k(k-1)}{2N_e(0)}\left(e^{\beta t_{k-1}} - e^{\beta t_k}\right)\right\}$$

Mutation in the Coalescent:

A mutational model for the Coalescent Process is derived by thinking once again in discrete generations and then making a continuous time approximation. Let μ denote the probability of the offspring of a gene, from one generation to the next, is a mutant. Let γ_r be the total number of mutations accumulated in one gene line of descent after *r* generations. Under the assumption of independence across lineages, this number of mutations can be modeled with a binomial distribution with probability μ and total number of trials *r*. Denote S_2 the number of mutations separating two individuals. Conditional on the time U_2 (in discrete generations) until these two individuals find their most recent common ancestor, $(S_2|U_2 = u_2) \sim \mathrm{Binom}(u_{2'}\mu)$ and recalling that $E[U_2] = N_e$ it follows that $E(S_2) = E[E(S_2|U_2)] = E[2U_2\mu] = 2N_e\mu = \theta$. Using the same time-scale change defined above and replacing *r* with $N_e t$ the binomial pmf of γ_r becomes

$$\Pr(\gamma_{Nt} = j) = \binom{Nt}{j}\left(\frac{\theta}{2N_e}\right)^j\left(1 - \frac{\theta}{2N_e}\right)^{Nt-j} \to \frac{1}{j!}\left(\frac{\theta t}{2}\right)^j e^{-\theta t/2}$$

as $N_e \to \infty$. Thus, mutations in the Coalescent are simply modeled with a Poisson process with rate $\theta t \,/\, 2$. Critical for this derivation is the conditioning step, and the integration (i.e. calculation of the expected value or average) over all the possible genealogy lengths separating two individuals. The same integration is needed to compute the overall likelihood functions.

Likelihood function under the Coalescent with mutations:

The reader familiar with hierarchical or "state-space models" in biology, will recognize the Coalescent process with mutation is indeed a hierarchical stochastic model. Such models allow researchers to incorporate variability in parameters that otherwise might be unrealistically treated as fixed. In addition, these models allow the incorporation of multiple layers of process and/or observation variability. Until recently, computational difficulties rendered likelihood inference for these models unfeasible, or plainly unreliable. For all but the simplest models, the likelihood function is written as a multi-dimensional integral. Here we solve this integration problem using Data Cloning (DC), which is an efficient and extensively tested computational algorithm to find the Maximum Likelihood (ML; 20, 21, 22, 51-56). The DC theorem allows one to apply a typical Bayesian posterior calculation and MCMC sampling to a number c of copies (clones) of the data (52). When c is large, the sample mean vector of the resulting simulated posterior distribution corresponds to the ML estimates of the parameters. Furthermore, the sample variance-covariance matrix of the posterior, multiplied by c, provides estimates of the variances and co-variances of these ML estimates (the inverse of the observed Fisher's information matrix). Ponciano et al (22) extended this estimation methodology to a complete inferential approach by proving and demonstrating how DC for hierarchical models can be easily extended to carry model selection, Likelihood Ratio Tests (LRTs) and computing profile likelihood intervals with much better coverage than the Wald confidence intervals for small sample sizes. This DC methodology is what we use here. We refer the interested reader to Ponciano et al. (20) who show step by step the explicit DC calculations for an analytically tractable example. We favored this methodology because, unlike any available Bayesian software to work with the Coalescent process, we can (and did) explicitly and efficiently assess the identifiability and estimability of the model parameters. This assessment is the greatest advantage of using DC for hierarchical

models vs. conforming to a Bayesian estimation methodology. Here again, we refer the reader to Ponciano et al. (20) for explicit and extensive accounts of such assessment. In the results section we illustrate the assessment of parameter identifiability using the data coming from one zebra.

With a sample of size *n* a total of S_n segregating sites are observed, the likelihood function is written as the Poisson probability with S_n variants emerging along the genealogy, averaged over all possible genealogies. The joint distribution of the inter-Coalescent times $u_i, i = n, n-1, \ldots, 2$ (see Figure 1) is simply the product of their pdfs $f(u_k) = Pr(u_k|t_k)$. For the constant N_e population model, this product is $\underline{u}$

$$f(u_2)f(u_3)\ldots f(u_n) = f(\underline{u}) = \prod_{k=2}^{n} \frac{k(k-1)}{2} e^{\left\{\frac{k(k-1)}{2}u_k\right\}}$$

whereas for the exponential model where it is assumed the population decays exponentially from the present to the past according to the model $N_e(t) = N_e(0)e^{-\beta t}$, $\left(e^{\beta t_{k-1}} - e^{\beta t_k}\right)$

$$f(u_2)f(u_3)\ldots f(u_n) = f(\underline{u}) = \prod_{k=2}^{n} \frac{k(k-1)e^{\beta t_{k-1}}}{2N_e(0)} \times \exp^{\left\{-\frac{k(k-1)}{2N_e(0)\beta}\left(e^{\beta t_{k-1}} - e^{\beta t_k}\right)\right\}}$$

.

Since along a branch of length *u* of the genealogy, the number of mutations is distributed Poisson with mean $\frac{\theta u}{2}$ for the constant effective population size model, given a particular genealogy (i.e.,

given an particular set of values of $u_n, u_{n-1}, \ldots, u_2$), the conditional distribution of the total number of mutations $S_n \mid (u_n, u_{n-1}, \ldots, u_2)$ along this genealogy is going to be Poisson distributed with mean $\frac{\theta L}{2}$ where

$L = \sum_{i=2}^{n} iu_i$ is the total length of a given genealogical tree (SI Appendix see Fig 1). That is

$$\Pr(S_n = s|(u_n, u_{n-1}, \ldots, u_2) = \frac{e^{-\frac{\theta L}{2}}(\theta L/2)^s}{s!}$$

For the exponential growth model the value of θ changes over time according to $\theta(t) = \theta_0 e^{-\beta t}$ and we arbitrarily assume such changes only occur at the Coalescent events and therefore,

$$\Pr(S_n = s|(u_n, u_{n-1}, \ldots, u_2) = \frac{e^{-\sum_{i=2}^{n} \frac{\theta_{i-1} i u_i}{2}} \left(\sum_{i=2}^{n} \frac{\theta_{i-1} i u_i}{2}\right)^s}{s!}.$$

Averaging these Poisson probabilities over all the possible genealogy lengths gives us the likelihood function as

$$\Pr(S_n = s) = \int \ldots \int \Pr(S_n = n|u_2, u_3 \ldots u_n) f(u_2) f(u_3) \ldots f(u_n) du_2 \ldots du_n$$

Both likelihood functions were maximized in JAGS (57) using the DC methodology. Our computer code is available in the appendix. After maximizing the likelihood, we used the methodology in Ponciano et al. (22) to compute the ML estimates of the latent variables $u_2, u_2, \ldots, u_n$ and of their sum, which is the TMRCA. We also used Ponciano et al.'s (22) DC likelihood ratio test and model selection tools to test the goodness of fit of the exponential vis-à-vis the constant population size model for the data in all zebra. For the laboratory data, we assumed the constant population size model (58). Joyce et al. (58) demonstrated the overall dynamics of a serial passage experiment with plasmid carrying and plasmid-free bacteria mirrored the dynamics during a single day because bacteria were grown approximately to the same total from one cycle to the next of the experiment. Under these conditions, the bacterial

dynamics could be accurately predicted (59) and estimated by assuming a constant bacterial population size at the end of each cycle. The alternative would be to fit a Coalescent model with as many bottlenecks as serial passage transfers, which is beyond the scope of this work. The laboratory constant population size assumption allowed us to estimate the laboratory N_e directly from the Coalescent time scaling and the known number of elapsed generations throughout the experiment (214, at 6 generations per day). Since our Coalescent model fitting gave us the ML estimate of the TMRCA and one unit of the Coalescent time corresponds to N_e discrete generations, we simply obtained our N_e estimate as 214/TMRCA. Since our model fitting also gives us an independent estimate of Θ for the laboratory, we could solve for the per generation mutation rate $\mu = 0.002$.

Finally, the value of $N_e(t)$ in the above likelihood can be arbitrarily substituted by $\Theta(t)$ without affecting the maximum location in parameter space (60-62). After all, both quantities are proportional to each other. After maximization, whenever we fitted the constant population size we accomplished the transformation from values of Θ to values of N_e by dividing by twice the laboratory rate mutation rate per generation μ. Recalling one unit of Coalescent time corresponds to N_e generations for this simple model and knowing the number of generations per day is approximately six, we then transformed the ML estimate of the TMRCA to days and took this value as the estimate of the retrospective number of days from death to infection. For the exponential model, the transformation from Coalescent time to generations was accomplished by solving the question: How many discrete generations j does it take to traverse τ units of exponentially decaying Coalescent time, starting from the present to the past?

Suppose the population size j generations back into the past, corresponding to τ Coalescent time units is $N(j)$. Because the amount of Coalescent time traversed from generation

i to i+1 back in the past is $\frac{1}{N_e(i)'}$ then during j generations, the total amount of Coalescent time τ is given by

$$\tau = g(j) = \sum_{j=1}^{j} \frac{1}{N_e(i)}$$

.

Having an estimate of τ (which for us will be the TMRCA) all we did was to solve for j in the above equation, by using the exponential growth model $N_e(t) = N_e(0)\mathrm{e}^{-\beta t}$ and the integral approximation

$$\sum_{j=1}^{j} \frac{1}{N_e(i)} \approx \int_0^j \frac{1}{N_e(s)} ds = \frac{1}{N_e(0)\beta}\left(e^{\beta j} - 1\right)$$

. Accordingly, $j = \frac{ln(N_e(0)\beta\tau+1)}{\beta}$.

For both models, we transformed the time to most recent common ancestor from Coalescent time units to real time units assuming two possible values of N_e. First, we estimated N_e using the mutation rate estimated from the laboratory experiment and the ML estimate of Θ for each zebra and either the constant population size or exponential population growth models. For the exponential population size model, we then estimated the initial N_e when each zebra was infected using the ML estimates of β in each zebra.

Data Availability

All data and detailed methods are available upon request to WCT or NCS. This includes detailed protocols, data (cfu counts and timetables for the transfer exp., photos of sampled colonies for

the mutation rate exp. genotype data including raw fragment size data, etc.) and code for coalescence modeling.

Acknowledgements:

We thank the Ministry of Environment and Tourism in Namibia for permission to conduct research in Etosha National Park, and we are grateful to the scientific staff and managers at the Etosha Ecological Institute for logistical support and assistance. We thank Zoe Barandongo, Claudine Cloete, and Clemens Naomob for laboratory assistance.

Funding was provided by NSF OISE-1103054 and NSF DEB-1816161 (to WCT), and RCN 225031/E31 (to NCS).

References:

1. Koch R (1876) The etiology of anthrax, based on the life history of Bacillus anthracis. *Beiträge zur Biologie der Pflanzen* 2(2):277-310.
2. Lee A*, et al.* (1997) A standardized mouse model of Helicobacter pylori infection: Introducing the Sydney strain. *Gastroenterology* 112(4):1386-1397.
3. Santos RL*, et al.* (2001) Animal models of Salmonella infections: enteritis versus typhoid fever. *Microbes and Infection* 3(14):1335-1344.
4. Jirtle RL & Skinner MK (2007) Environmental epigenomics and disease susceptibility. *Nature Reviews Genetics* 8:253.
5. Hugh-Jones M & Blackburn J (2009) The ecology of Bacillus anthracis. *Molecular Aspects of Medicine* 30(6):356-367.
6. Hugh-Jones ME & de Vos V (2002) Anthrax and wildlife. *Revue scientifique et technique (International Office of Epizootics)* 21(2):359-383.
7. Blackburn JK, Van Ert M, Mullins JC, Hadfield TL, & Hugh-Jones ME (2014) The necrophagous fly anthrax transmission pathway: empirical and genetic evidence from wildlife epizootics. *Vector-Borne and Zoonotic Diseases* 14(8):576-583.
8. Turner WC*, et al.* (2016) Lethal exposure: An integrated approach to pathogen transmission via environmental reservoirs. *Scientific Reports* 6:27311.
9. Glomski IJ, Piris-Gimenez A, Huerre M, Mock M, & Goossens PL (2007) Primary Involvement of Pharynx and Peyer's Patch in Inhalational and Intestinal Anthrax. *PLOS Pathogens* 3(6):e76.
10. Thompson BM, Waller LN, Fox KF, Fox A, & Stewart GC (2007) The BclB Glycoprotein of Bacillus anthracis Is Involved in Exosporium Integrity. *Journal of Bacteriology* 189(18):6704-6713.

11. Gu C, Jenkins SA, Xue Q, & Xu Y (2012) Activation of the Classical Complement Pathway by *Bacillus anthracis* Is the Primary Mechanism for Spore Phagocytosis and Involves the Spore Surface Protein BclA. *The Journal of Immunology* 188(9):4421-4431.
12. Barandongo ZR, Mfune JKE, & Turner WC (2018) DUST-BATHING BEHAVIORS OF AFRICAN HERBIVORES AND THE POTENTIAL RISK OF INHALATIONAL ANTHRAX. *Journal of Wildlife Diseases* 54(1):34-44.
13. Turner WC*, et al.* (2014) Fatal attraction: vegetation responses to nutrient inputs attract herbivores to infectious anthrax carcass sites. *Proceedings of the Royal Society B: Biological Sciences* 281(1795).
14. Cizauskas CA, Bellan SE, Turner WC, Vance RE, & Getz WM (2014) Frequent and seasonally variable sublethal anthrax infections are accompanied by short-lived immunity in an endemic system. *J Anim Ecol* 83.
15. Bellan SE, Gimenez O, Choquet R, & Getz WM (2013) A hierarchical distance sampling approach to estimating mortality rates from opportunistic carcass surveillance data. *Methods in Ecology and Evolution* 4(4):361-369.
16. Kingman JFC (1982) The Coalescent. *Stochastic processes and their applications* 13(3):235 - 248.
17. Ewens W & Joyce P (2009) Mathematical Population Genetics, Introduction to the Stochastic Theory. (Center for Research in Mathematics, CIMAT, Guanajuato, Mexico, Lecture Notes of the Summer School in Probability and Statistics).
18. Ponciano JM (2018) A parametric interpretation of Bayesian Nonparametric Inference from Gene Genealogies: Linking ecological, population genetics and evolutionary processes. *Theor Popul Biol* 122:128-136.
19. Palacios JA & Minin VN (2013) Gaussian Process-Based Bayesian Nonparametric Inference of Population Size Trajectories from Gene Genealogies. *Biometrics* 69(1):8-18.
20. Ponciano JM, Burleigh JG, Braun EL, & Taper ML (2012) Assessing Parameter Identifiability in Phylogenetic Models Using Data Cloning. *Syst Biol* 61(6):955-972.
21. Lele SR, Dennis B, & Lutscher F (2007) Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecol Lett* 10(7):551-563.
22. Ponciano JM, Taper ML, Dennis B, & Lele SR (2009) Hierarchical models in ecology: confidence intervals, hypothesis testing, and model selection using data cloning. *Ecology* 90(2):356-362.
23. Anonymous (2008) Anthrax in Humans and Animals. ed Turnbull P (World Health Organization), 4 Ed.
24. Dialdestoro K*, et al.* (2016) Coalescent Inference Using Serially Sampled, High-Throughput Sequencing Data from Intrahost HIV Infection. *Genetics* 202(4):1449-1472.
25. Achtman M (2008) Evolution, Population Structure, and Phylogeography of Genetically Monomorphic Bacterial Pathogens. *Annual Review of Microbiology* 62(1):53-70.
26. Stratilo CW & Bader DE (2012) Genetic Diversity among Bacillus anthracis Soil Isolates at Fine Geographic Scales. *Applied and Environmental Microbiology* 78(18):6433-6437.
27. Kenefic L*, et al.* (2008) A high resolution four-locus multiplex single nucleotide repeat (SNR) genotyping system in Bacillus anthracis. *J Microbiol Meth* 73(3):269-272.
28. Braun P*, et al.* (2015) Microevolution of Anthrax from a Young Ancestor (M.A.Y.A.) Suggests a Soil-Borne Life Cycle of Bacillus anthracis. *PLOS ONE* 10(8):e0135346.
29. Turell MJ & Knudson GB (1987) Mechanical transmission of Bacillus anthracis by stable flies (Stomoxys calcitrans) and mosquitoes (Aedes aegypti and Aedes taeniorhynchus). *Infection and Immunity* 55(8):1859-1861.
30. Basson L*, et al.* (2018) Blowflies as vectors of Bacillus anthracis in the Kruger National Park. *2018* 60(1).

31. Beyer W & Turnbull PCB (2009) Anthrax in animals. *Molecular Aspects of Medicine* 30(6):481-489.
32. Cizauskas CA*, et al.* (2014) Gastrointestinal helminths may affect host susceptibility to anthrax through seasonal immune trade-offs. *BMC Ecology* 14(1):27.
33. Dixon TC, Fadl AA, Koehler TM, Swanson JA, & Hanna PC (2000) Early Bacillus anthracis–macrophage interactions: intracellular survival and escape. *Cellular Microbiology* 2(6):453-463.
34. Hu H, Emerson J, & Aronson AI (2007) Factors involved in the germination and inactivation of Bacillus anthracis spores in murine primary macrophages. *FEMS Microbiology Letters* 272(2):245-250.
35. Lowe DE, Ernst SMC, Zito C, Ya J, & Glomski IJ (2013) Bacillus anthracis Has Two Independent Bottlenecks That Are Dependent on the Portal of Entry in an Intranasal Model of Inhalational Infection. *Infection and Immunity* 81(12):4408-4420.
36. de Vos V (1990) The ecology of anthrax in the Kruger National Park, South Africa. *Salisbury Medical Bulletin* 68S:19-23.
37. Turner WC*, et al.* (2013) Soil ingestion, nutrition and the seasonality of anthrax in herbivores of Etosha National Park. *Ecosphere* 4(1):art13.
38. Lindeque PM & Turnbull PC (1994) Ecology and epidemiology of anthrax in the Etosha National Park, Namibia. *The Onderstepoort journal of veterinary research* 61(1):71-83.
39. Beyer W*, et al.* (2012) Distribution and Molecular Evolution of *Bacillus anthracis* Genotypes in Namibia. *PLoS Negl Trop Dis* 6(3):e1534.
40. Van Ert MN*, et al.* (2007) Global Genetic Population Structure of *Bacillus anthracis*. *PLoS ONE* 2(5):e461.
41. Bruce SA, Schiraldi NJ, Kamath PL, Easterday WR, & Turner WC (In Press) A classification framework for Bacillus anthracis defined by global genomic structure. *Evolutionary Applications*.
42. Lista F*, et al.* (2006) Genotyping of Bacillus anthracis strains based on automated capillary 25-loci multiple locus variable-number tandem repeats analysis. *BMC Microbiol* 6(1):33.
43. Tavare S, Balding DJ, Griffiths RC, & Donnelly P (1997) Inferring coalescence times from DNA sequence data. *Genetics* 145(2):505-518.
44. Watterson GA (1975) Number of Segregating Sites in Genetic Models without Recombination. *Theor Popul Biol* 7(2):256-276.
45. Ward RH, Frazier BL, Dewjager K, & Paabo S (1991) Extensive Mitochondrial Diversity within a Single Amerindian Tribe. *P Natl Acad Sci USA* 88(19):8720-8724.
46. Wakeley J & Sargsyan O (2009) Extensions of the Coalescent Effective Population Size. *Genetics* 181(1):341-345.
47. Rambaut A*, et al.* (2008) The genomic and epidemiological dynamics of human influenza A virus. *Nature* 453(7195):615-U612.
48. Frost SDW & Volz EM (2010) Viral phylodynamics and the search for an 'effective number of infections'. *Philos T R Soc B* 365(1548):1879-1890.
49. Felsenstein J (2004) *Inferring phylogenies* (Sunderland: Sinauer associates).
50. Griffiths RC & Tavare S (1994) Ancestral Inference in Population-Genetics. *Stat Sci* 9(3):307-319.
51. Bolker BM*, et al.* (2009) Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol Evol* 24(3):127-135.
52. Lele SR, Nadeem K, & Schmuland B (2010) Estimability and Likelihood Inference for Generalized Linear Mixed Models Using Data Cloning. *J Am Stat Assoc* 105(492):1617-1625.
53. Solymos P (2010) dclone: Data Cloning in R. *R J* 2(2):29-37.
54. Baghishani H & Mohammadzadeh M (2011) A data cloning algorithm for computing maximum likelihood estimates in spatial generalized linear mixed models. *Comput Stat Data An* 55(4):1748-1759.

55. Campbell D & Lele S (2014) An ANOVA test for parameter estimability using data cloning with application to statistical inference for dynamic systems. *Comput Stat Data An* 70:257-267.
56. Gomez JP, Robinson SK, Blackburn JK, & Ponciano JM (2016) An efficient extension of N-mixture models for multi-species abundance estimation. *BioRxiv*:073577.
57. Plummer M (2003) JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the third international workshop on distributed statistical computing* 124:125.
58. Joyce P*, et al.* (2005) Modeling the impact of periodic bottlenecks, unidirectional mutation, and observational error in experimental evolution. *J Math Biol* 50(6):645-662.
59. De Gelder L*, et al.* (2004) Combining mathematical models and statistical methods to understand and predict the dynamics of antibiotic-sensitive mutants in a population of resistant bacteria during experimental evolution. *Genetics* 168(3):1131-1144.
60. Pybus OG, Rambaut A, & Harvey PH (2000) An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* 155(3):1429-1437.
61. Minin VN, Bloomquist EW, & Suchard MA (2008) Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol Biol Evol* 25(7):1459-1471.
62. Paradis E & Paradis ME (2015) Package 'coalescentMCMC'. . *Biometrika* 57:97 - 109.

Figure 1. Histograms of the Time to Most Recent Common Ancestor (TMRCA) for 11 Zebra carcasses plotted for 50000 samples of the posterior distribution given the likelihood of the constant population size (black) exponential population growth (grey) models. According to data cloning theory (DC), the Maximum Likelihood estimate of TMRCA (red vertical line) is given by the mean of this 50000 samples. The estimates have been rescaled to represent time in days and not Coalescent time.

Figure 2: Illustration of population dynamics of *B. anthracis* through infection-transmission cycles for log(*N*) *B. anthracis* population [shaded yellow] over time [split into days and years]). **A) *Ingestion:*** ungulates grazing at carcass sites ingest a portion of the spores present along with forage and soil, creating a bottleneck. **B**) ***crossing epithelium***: after ingestion only a portion of the ingested cells cross the epithelium, starting the infection. **C**) ***climax population:*** the population climax, near the time of death. **D**) ***decay:*** the process of spore decay begins *post-mortem*.

	Constant			Exponential				
Zebra No	**θ**	**N_e**	**TMRCA**	**θ_0**	**β**	**N_e(1)**	**TMRCA**	**CFU**
1	1.05	286.16	88.14	1.9(1.4,2.63)	0.69(0.2,1.18)	215.94(120.18,528.12)	1.47(0.86,4.35)	1.71 (0.95, 4.17)
2	1.08	294.08	89.41	1.92(1.39,2.7)	0.74(0.23,1.23)	212.13(116.52,514.11)	1.39(0.83,3.94)	1.68 (0.92, 4.06)
3	0.79	215.5	67.18	1.88(1.38,2.62)	0.57(0.1,1.08)	232.67(118.23,634.15)	1.74(0.94,7.33)	1.84 (0.93, 5.01)
5	0.52	142.3	44.13	1.88(1.38,2.61)	0.36(0.1,0.89)	295.38(126.63,636.11)	2.61(1.13,7.22)	2.34 (1.00, 5.03)
7	1.09	297.81	90.85	1.92(1.38,2.72)	0.76(0.25,1.26)	209.15(114.66,508.35)	1.36(0.81,3.75)	1.65 (0.91, 4.02)
8	1.03	280.55	82.05	2.02(1.27,3.33)	0.96(0.29,1.6)	193.25(88.07,621.59)	1.11(0.64,3.37)	1.53 (0.70, 4.91)
9	0.6	163.24	49.7	1.91(1.31,2.88)	0.58(0.1,1.18)	235.22(101.56,708.49)	1.74(0.86,7.23)	1.86 (0.80, 5.60)
13	1.1	301.08	91.46	1.92(1.38,2.75)	0.77(0.25,1.28)	208.01(112.63,509.58)	1.33(0.8,3.69)	1.64 (0.89, 4.03)
14	0.52	142.3	44.13	1.88(1.38,2.61)	0.36(0.1,0.89)	295.38(126.63,636.11)	2.61(1.13,7.22)	2.34 (1.00, 5.03)
17	0.53	145.22	45.06	1.89(1.38,2.66)	0.45(0.1,1)	260.27(116.15,650.11)	2.15(1.01,7.21)	2.06 (0.92, 5.14)
19	0.28	77.24	24.26	2.42(1.67,3.64)	1.62(0.63,2.47)	118.09(52.91,375.73)	0.73(0.47,1.81)	0.93 (0.42, 2.97)

Table 1. Parameter estimates for both constant size and exponential population growth models. θ is the average number of mutations that separates two genes under the Coalescent process. It is defined as twice the effective population size N_e times the mutation rate μ. This number remains the same under the constant effective population size model. Under the exponential population growth model, the zebra's *B. anthracis* population value of θ at the moment of death is θ_0 and the effective population size changes (from present to past) according to the exponential function $N_e(t) = N_e(0)e^{-\beta t}$, where β is the exponential rate parameter and $N_e(0) = \frac{\theta_0}{2\mu}$. Accordingly, $N_e(1)$ represents the effective population size of *B. anthracis* in each zebra at moment of infection using the experiment's estimated mutation rate (see full model and statistical analyses description in Methods). Confidence intervals are calculated only for the exponential population

growth model since it was the best fit to the data. TMRCA is the estimated Time to Most Recent Common Ancestor expressed in days assuming a mutation rate of 0.002.