MICROBIAL GENOMICS

RESEARCH ARTICLE

Bruce *et al.*, *Microbial Genomics* 2021;7:000616

DOI 10.1099/mgen.0.000616

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

The roles of antimicrobial resistance, phage diversity, isolation source and selection in shaping the genomic architecture of *Bacillus anthracis*

Spencer A. Bruce[1,*], Yen-Hua Huang[2], Pauline L. Kamath[3], Henriette van Heerden[4] and Wendy C. Turner[5]

Abstract

Bacillus anthracis, the causative agent of anthrax disease, is a worldwide threat to livestock, wildlife and public health. While analyses of genetic data from across the globe have increased our understanding of this bacterium's population genomic structure, the influence of selective pressures on this successful pathogen is not well understood. In this study, we investigate the effects of antimicrobial resistance, phage diversity, geography and isolation source in shaping population genomic structure. We also identify a suite of candidate genes potentially under selection, driving patterns of diversity across 356 globally extant *B. anthracis* genomes. We report ten antimicrobial resistance genes and 11 different prophage sequences, resulting in the first large-scale documentation of these genetic anomalies for this pathogen. Results of random forest classification suggest genomic structure may be driven by a combination of antimicrobial resistance, geography and isolation source, specific to the population cluster examined. We found strong evidence that a recombination event linked to a gene involved in protein synthesis may be responsible for phenotypic differences between comparatively disparate populations. We also offer a list of genes for further examination of *B. anthracis* evolution, based on high-impact single nucleotide polymorphisms (SNPs) and clustered mutations. The information presented here sheds new light on the factors driving genomic structure in this notorious pathogen and may act as a road map for future studies aimed at understanding functional differences in terms of *B. anthracis* biogeography, virulence and evolution.

DATA SUMMARY

(1) All NCBI accession numbers related to sequence reads and bioproject data used in this study are listed in File S1.

(2) All supplementary material can be found at https://doi.org/10.6084/m9.figshare.14485140.v1

(3) The R script used for the random forest classification and code used for identifying clustered mutations can be found at the following GitHub repository: https://github.com/spencer411/B_anthracis_adaptation.

INTRODUCTION

For pathogens, consideration of intraspecific variation is central to understanding the evolution of virulence and genotypic persistence [1–3]. Phenotypic and genetic variation in a population may influence ecological composition and function, leading to increased or decreased evolutionary capacity under altered habitat regimes [4, 5]. Incorporating intraspecific diversity into effective management strategies demands the identification of factors influencing ecological plasticity and reproductive success [6, 7]. A wide range of genomic analyses have revealed genetic anomalies supporting ecologically variable phenotypes, suggesting a consequential

Received 23 February 2021; Accepted 23 May 2021; Published 17 August 2021

Author affiliations: [1]Department of Biological Sciences, University at Albany – State University of New York, Albany, NY 12222, USA; [2]Wisconsin Cooperative Wildlife Research Unit, Department of Forest and Wildlife Ecology, University of Wisconsin-Madison, Madison, WI, USA; [3]School of Food and Agriculture, University of Maine, Orono, ME 04469, USA; [4]Department of Veterinary Tropical Diseases, University of Pretoria, Onderstepoort, South Africa; [5]U.S. Geological Survey, Wisconsin Cooperative Wildlife Research Unit, Department of Forest and Wildlife Ecology, University of Wisconsin-Madison, Madison, WI, USA.

***Correspondence:** Spencer A. Bruce, sbruce@albany.edu

Keywords: AMR; *Bacillus anthracis*; bacterial genomics; natural selection.

Abbreviations: AMR, antimicrobial resistance; OOB, out-of-bag; PA, protective antigen; RF, random forest; RGI, resistance gene identifier; SNP, single nucleotide polymorphism; TPR, tetratricopeptide repeat.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. One supplementary figure, eight supplementary tables and one supplementary information file are available with the online version of this article role for genomic architecture in driving intraspecific heterogeneity [8, 9]. For example, single nucelotide polymorphisms (SNPs) may facilitate the evolution of new phenotypes through the formation of novel proteins in regions that code for spore formation in *Bacillus anthracis* or virulence factors in *Clostridium difficile* [10, 11]. By evaluating the relationship between whole genome architecture and ecologically relevant sequence variation, we can gain a detailed understanding of how biocomplexity drives genomic structure in pathogenic bacteria [12, 13]. Nevertheless, the relationship between genomic variation and adaptive relevance remains largely unknown for the vast majority of pathogenic species [14, 15].

B. anthracis has been extensively studied given its ability to cause anthrax, a disease that can be fatal to wildlife, livestock and humans [16]. Studies that have examined the integration of bacteriophage DNA into the *B. anthracis* genome have suggested that these sequences may influence gene expression, potentially driving increased sporulation and observable phenotypic differences [17, 18]. In addition, antimicrobial resistance (AMR) has recently garnered a great deal of attention given the wide range of antibiotics administered to both humans and livestock throughout the world, driving selective resistance in a myriad of bacteria including *B. anthracis* [19, 20]. Therefore, when examining *B. anthracis* genomic architecture in light of selection, the use of classification methodologies that incorporate potentially ecologically relevant differences in phage diversity and AMR may shed light on the drivers of modern population genomic structure in this species. This in turn will allow us to better forecast what genomic clusters or clades may pose the greatest risk of disease emergence and re-emergence in animals and humans [21]. Nevertheless, it should be noted that the detection of an AMR gene does not always translate to conferred resistance [22].

Individual and regional genetic diversity that differentiates *B. anthracis* populations by SNP architecture has been identified on a global scale [23–26]. Recent work has refined our understanding of population genomic structure for this species [27, 28]. Work by Sahl *et al*. sought to expand on the original *B. anthracis* classification system, and generated an SNP database used to characterize the branching structure of isolates based on 193 genomes [28]. More recent genomic analyses that comprise the largest global phylogeny of *B. anthracis* to date (356 genomes) has redefined *B. anthracis* population genomic structure, resulting in six primary clusters and 18 nested clades. This new classification system uses an intuitive, simplified naming system and allows for linkable, rapid classification [27]. Two of the major genotype clusters, cluster 1 (C Branch) and cluster 2 (B Branch), are vastly underrepresented in terms of prevalence and have been hypothesized to be less fit than the majority of *B. anthracis* specimens isolated and sequenced [26, 29, 30]. However, the link between genomic architecture, and the scarcity of these genotypes remains largely unexamined. In addition, some of the individual clades identified are geographically specific, whereas others seem to be widely distributed, raising numerous questions about what factors are driving evolutionary success in this species [26–28]. Understanding the relationships among spatial variation, population stability, and genomic architectural variation is particularly important for *B. anthracis*, as it is a major threat to wildlife, livestock and public health globally [31]. In this study, we explore genomic variation and selection in a global whole-genome dataset of *B. anthracis* isolates spanning 39 countries and six continents. In addition, we apply an ensemble machine learning method [random forest (RF)] to elucidate the ways in which isolation source, geography, phage diversity and AMR genes may be shaping genomic diversity and genotypic persistence. RF operates by constructing decision trees on various subsamples of the dataset, allowing for predictions regarding evolutionary potential at the population level.

Impact Statement

Understanding the drivers of pathogen genomic structure allows for targeted disease management based on factors contributing to virulence and host susceptibility. Despite the large range of published information on *B. anthracis* genetic structure, little work has been done to understand the factors shaping its global genetic constitution. The data presented here allow for the first large-scale accounting of antimicrobial resistance and phage sequence diversity for this species. These results suggest that antibiotic resistance genes and isolation source may be driving aspects of population structure and emphasize the importance of examining multiple factors dictating pathogen evolution and genotypic persistence.

METHODS

Whole genome mapping and assembly

The population genomic dataset used in these analyses was previously developed and published by Bruce *et al*. [27] consisting of 356 *B. anthracis* whole genomes collected from the NCBI sequence read archive ([27], File S1, available in the online version of this article). Each read pair was mapped to the fully annotated Ames Ancestor genome (accession AE017334.2), using the RedDog pipeline (https://github.com/katholt/RedDog). Mapped reads were then subjected to extensive post-processing to remove calls (a) found in regions with large 'inexact' repeats, (b) within prophage regions of the reference genome, (c) from regions that were found to be invariable in all but the outgroup, (d) from regions potentially resulting from recombination and (e) potentially related to stutter. Full details relating to methods for mapping, SNP calling and determination of population genomic structure can be found in Bruce *et al*. [27].

For the purpose of this study, the same trimmed sequence reads were also subjected to *de novo* assemblies using SPAdes version 3.13.0, a genome assembly algorithm specifically developed for single cell and multi-cell bacterial isolates [32]. *De novo* assemblies allow for the identification of unique sequences in each isolate not identifiable using the mapping method described above.

Identification of AMR genes and phage sequence variation

We screened each assembly for AMR genes employing The Resistance Gene Identifier (RGI) tool provided by the Comprehensive Antibiotic Resistance Database (CARD) [33]. RGI can be used to predict resistomes from protein or nucleotide data based on homology and SNP models. In addition to identifying AMR genes, we identified prophage sequences within the contigs of each assembled genome using the Phage Search Tool Enhanced Release (PHASTER) [34]. PHASTER is a web-based application that is designed to rapidly and accurately identify, annotate and graphically display prophage sequences within bacterial genomes or plasmids. The full phylogenetic tree of *B. anthracis* isolates from Bruce *et al.* [27] was then annotated using iTOL [35] with both phage sequence variation (scored as either intact, questionable or incomplete; see Table S1 for details), and presence (or absence) of AMR genes. AMR gene data were then plotted geographically to understand patterns of resistance on a global scale using Adobe Illustrator [36].

Classifying population genomic architecture using RF

To understand how various factors may be influencing the genomic architecture of *B. anthracis* we used an RF approach [37], incorporating AMR gene data, phage diversity data and isolate metadata (continent of isolation and source) accessed through the NCBI biosample database [38]. RF has gained increased attention over the past several decades given its ability to produce excellent classification results while also being computationally inexpensive [39, 40]. The RF classifier produces valid classifications using predictions derived from a group of decision trees and can also be used to select and rank those variables, allowing the user to successfully discriminate between the target classes [41]. RF was carried out using the R package randomForest [37] to construct a multitude of decision trees and determine the mean prediction of each individual tree pertaining to the six primary population clusters [27]. The R package SPM was then used to carry out a 5-fold cross-validation [42].

We first removed samples with missing values for independent variables, and additionally removed three variables (AMR gene *mph*L, and phage sequence Bacillus virus 1 and Bacillus phage PfEFR-5) which exhibited no variation across the dataset. The final dataset resulted in 20 independent variables (Table S2). We divided the dataset into a training dataset including 75% of the samples and a validation dataset including the other 25%. To determine Mtry and Ntree (number of variables and number of trees), we used a 5-fold cross-validation and grid search. To carry out 5-fold cross-validation, we randomly assigned each sample to one of five groups. For each pass of cross-validation, RF classifiers were trained with a test dataset of which one group was held out [43]. The model with the highest correct classification rate and Kappa index of the classification was selected for determining values of Mtry and Ntree. We used the best combination of the Mtry and Ntree for the final RF model. To assess the model fit of the RF we subjected the model to the validation dataset and estimated the accuracy. To determine the contribution of the variables to the classification in the model, the importance of variables was evaluated by the mean decrease in accuracy. The mean decrease in accuracy was computed with the difference between the out-of-bag (OOB) error (training observations not included in the bootstrap) from a dataset with the selected variable permuted and the OOB error from the original dataset [41].

Recombination, high-impact SNPs and candidate genes for selection

To determine how recombination may be influencing population genomic structure across our dataset, we first used the program Gubbins to iteratively identify loci containing elevated densities of base substitutions in the SNP dataset (prior to removal of recombinant sequences) [44].

To analyse selection in non-recombining regions, we analysed SNPs (post-removal of recombinant sequences) using the program SnpEff [45]. SnpEff annotates and predicts the effects of genetic variants on genes and proteins (such as amino acid changes). To assess 'high-impact' SNPs influencing population genomic structure, we compiled a list of SNPs that produce significant changes to protein structure in the *B. anthracis* chromosome and plasmids, specific to each primary cluster and groups of primary clusters, such as mutations that result in the gain of a stop codon, the loss of a start codon and splice region variants. We also looked for clustered SNPs across each of the aforementioned groups to identify genes that were possibly associated with selection using a modified version of the algorithm developed by Cui *et al.* [46], classifying genes that showed three or more mutations within a 2000 bp range, as well as genes that showed two or more SNPs within a 50 bp range. Clustered mutations have a low probability of occurring under a neutral substitution model, in which variations are assumed to be randomly distributed across the genome Zhou *et al.* [47]. Examining the ratio of non-synonymous to synonymous SNPs at the gene level was problematic given the clonal nature of *B. anthracis* and reduced variability at the level of the gene, and was therefore not included in our analysis. We then compiled a list of candidate genes for selection that were identified using both methods above. Finally, we examined differences in SNP variation across the *B. anthracis* virulence genes (in the plasmids), again using SNPeff to identify SNPs potentially leading to functional differences across the different population genomic clusters.

RESULTS

Global variation in AMR and phage diversity

We identified a total of ten AMR genes across the global collection of 356 *B. anthracis* genomes analysed (Fig. 1, Table 1). Additional information regarding the AMR genes and their frequency is provided in Table S3. A key linking the classification framework shown here to the previously established branch labels outlined by Sahl *et al.* [28] are provided in Fig. S1. Five AMR genes (*mph*L, *bla*1, *fos*B, *bla*2 and *vml*R)

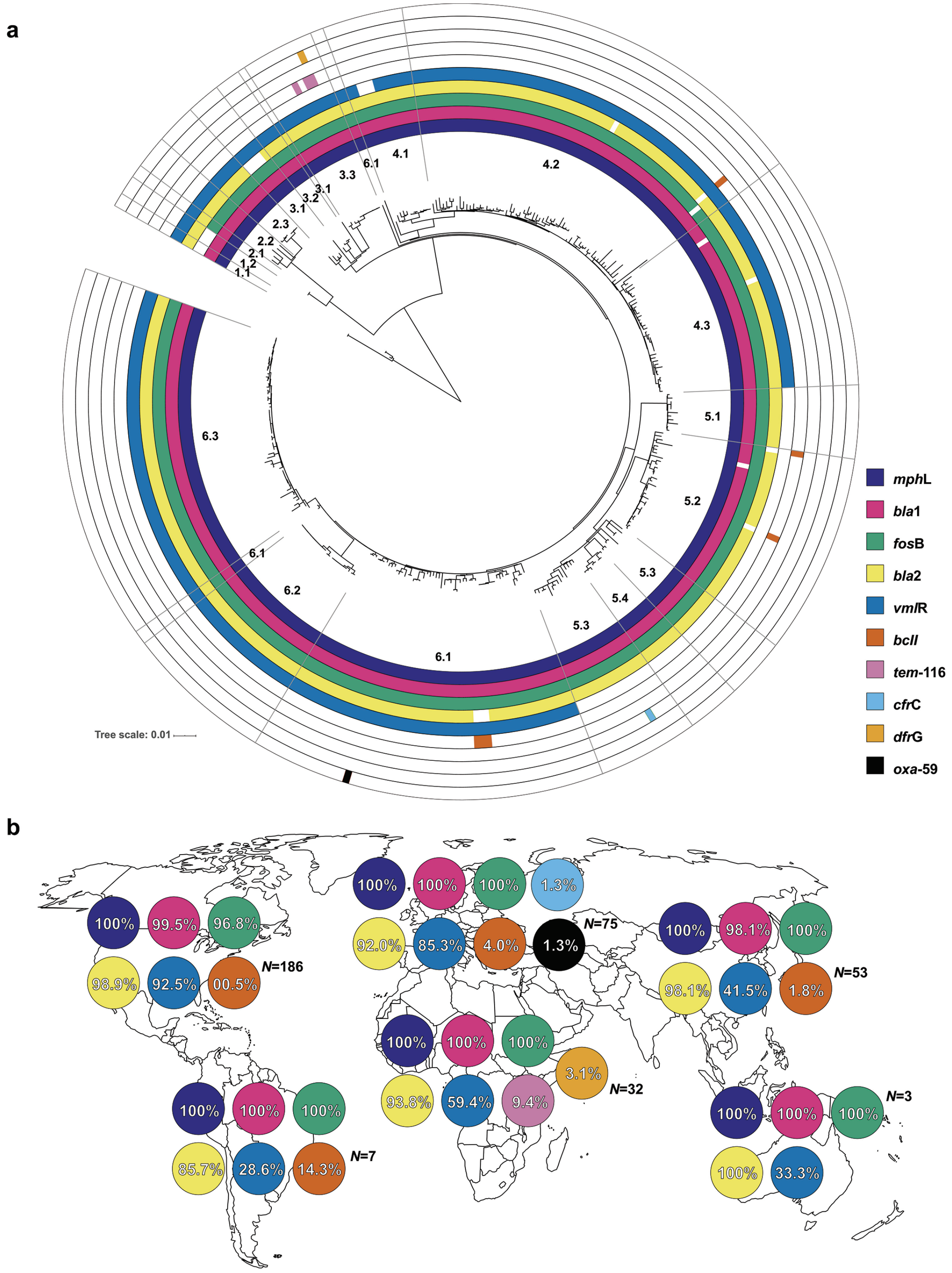

Fig. 1. AMR genes identified in the whole-chromosome tree of 356 global *B. anthracis* isolates (a). Primary clusters are divided into their numbered nested clades by grey lines. The key on the right lists the ten AMR genes identified. Outer rings reflect presence (colour) or absence (white) of each gene across all isolates in the phylogeny. A world map depicting the prevalence of AMR genes is depicted in (b). Each circle represents an AMR gene coloured according to the key and figure above. Percentages represent the total proportion of isolates from each continent where the respective AMR gene was identified (including North America, South America, Europe, Asia and Oceania). See Fig. S1 for a key to previously established classification schemes.

Table 1. AMR genes and their definitions from the Comprehensive Antibiotic Resistance Database (CARD)

Name	Resistance mechanism	Accession	Definition
*mph*L	Antibiotic inactivation	ARO:3003072	A chromosomally encoded macrolide phosphotransferase that inactivates macrolides such as erythromycin, clarithromycin, azithromycin
*bla*1	Antibiotic inactivation	ARO:3000090	A chromosomally encoded beta-lactamase that hydrolyses penicillins
*fos*B	Antibiotic inactivation	ARO:3000172	A thiol transferase that leads to fosfomycin resistance
*bla*2	Antibiotic inactivation	ARO:3004189	A chromosomally encoded beta-lactamase that has penicillin-, cephalosporin- and carbapenem-hydrolysing abilities
*vml*R	Antibiotic target protection	ARO:3004476	An ABC-F ATPase ribosomal protection protein shown to confer resistance to lincomycin and streptogramin A virginiamycin
bcII	Antibiotic inactivation	ARO:3002878	A zinc metallo-beta-lactamase that hydrolyses a large number of penicillins and cephalosporins
tem-116	Antibiotic inactivation	ARO:3000979	A broad-spectrum beta-lactamase found in many species of bacteria
*cfr*C	Antibiotic target alteration	ARO:3004146	A *cfr*-like 23S rRNA methyltransferase shown to confer resistance to linezolid and phenicol antibiotics, including florfenicol and chloramphenicol
*dfr*G	Antibiotic target replacement	ARO:3002868	A plasmid-encoded dihydrofolate reductase
oxa-59	Antibiotic inactivation	ARO:3001772	A beta-lactamase

were found across the majority of isolates tested. AMR gene *mph*L was identified in every isolate examined. AMR gene *bla*1 was absent in two unrelated isolates, one collected in South Carolina, USA [clade 4.3 (Vollum), NCBI Sequence Read Archive: SRR5811007], and the other collected in Morioka, Japan [clade 5.2 (Sterne), NCBI Sequence Read Archive: DRR128181]. AMR gene *fos*B was absent from a single isolate collected in South Carolina [clade 4.2 (Vollum), NCBI Sequence Read Archive: SRR5811063], and all isolates that comprise primary cluster 1 (C Branch) from the USA (*N*=5). The gene *bla*2 was absent from a number of other disparate samples (*N*=11) and was completely absent from all isolates that comprise clade 3.1 (Ancient A; *N*=4). AMR gene *vml*R was present in all isolates with the exception of a handful of closely related isolates collected in the USA between 1956 and 1978 from clade 4.1 (Vollum, *N*=3), as well as all isolates that comprise primary cluster 5 (V770, Ames, Sterne, Aust94; *N*=72). All other AMR genes were far rarer. AMR gene *bcII* was present in only six samples, including one isolate from Alabama [clade 4.2 (Vollum), NCBI Sequence Read Archive: SRR1739961], one isolate from Akita, Japan [clade 5.2 (Sterne), NCBI Sequence Read Archive: DRR128182], one isolate from Argentina [clade 5.2 (Sterne), NCBI Sequence Read Archive: SRR5810989], and three isolates from Albania [clade 6.1 (TEABr008/011), NCBI Sequence Read Archive: SRR2968139, SRR2968140 and SRR2968213]. AMR gene *tem*-116 was present in only three isolates, all collected in Zambia between 2012 and 2013 [clade 3.3 (Ancient A), NCBI Sequence Read Archive: DRR014736, DRR014737 and DRR125655]. AMR gene *cfr*C was present in a single isolate from Germany [clade 5.3 (Aust94), NCBI Sequence Read Archive: SRR2968155], *dfr*G was present in a single isolate from Zambia [clade 3.3 (Ancient A), NCBI Sequence Read Archive: DRR125655] and *oxa*-59 was present in a single isolate from Italy [clade 6.1 (TEABr008/011), NCBI Sequence Read Archive: SRR2968209].

In addition to AMR genes, we also identified 11 prophage sequences across our global dataset (Fig. 2, Table S4). Prophage sequences were scored as intact, questionable or incomplete. Criteria related to this categorization can be found in Table S1. Additional information regarding the phage sequences and their lineages is provided in Table S5. Bacillus virus 1, Bacillus phage PfEFR-5 and Staphylococcus phage vB_SepS_SEP9 sequences were detected across all of our samples. Bacillus virus 1 was determined to be intact in all isolates examined. Bacillus phage PfEFR-5 was determined to be questionable across most isolates, but incomplete for all isolates comprising primary cluster 1 (C Branch, *N*=5), while Staphylococcus phage vB_SepS_SEP9 was determined to be questionable across all isolates, but incomplete for all isolates comprising primary clusters 1 and 2 (C and B Branches; *N*=18). The eight remaining prophage sequences were scattered in comparatively minimal amounts across the global dataset, with the exception of Bacillus phage phBC6A52 which was intact in a large number of the isolates examined (*N*=91), with seemingly no link to relatedness or geography among isolates.

Explaining global genomic clusters with RF

The RF model was trained using 5-fold cross-validation with a training dataset. The best model parameter (where Ntree and Mtry equalled 400 and 9 respectively) produced a cross-correlation rate that showed a high value of 83.6, while kappa equalled 0.764. Variables examined include presence of AMR genes and phage sequences, as well as sample source (details provided in Table S2). With this combination of Ntree and Mtry, the OOB error based on

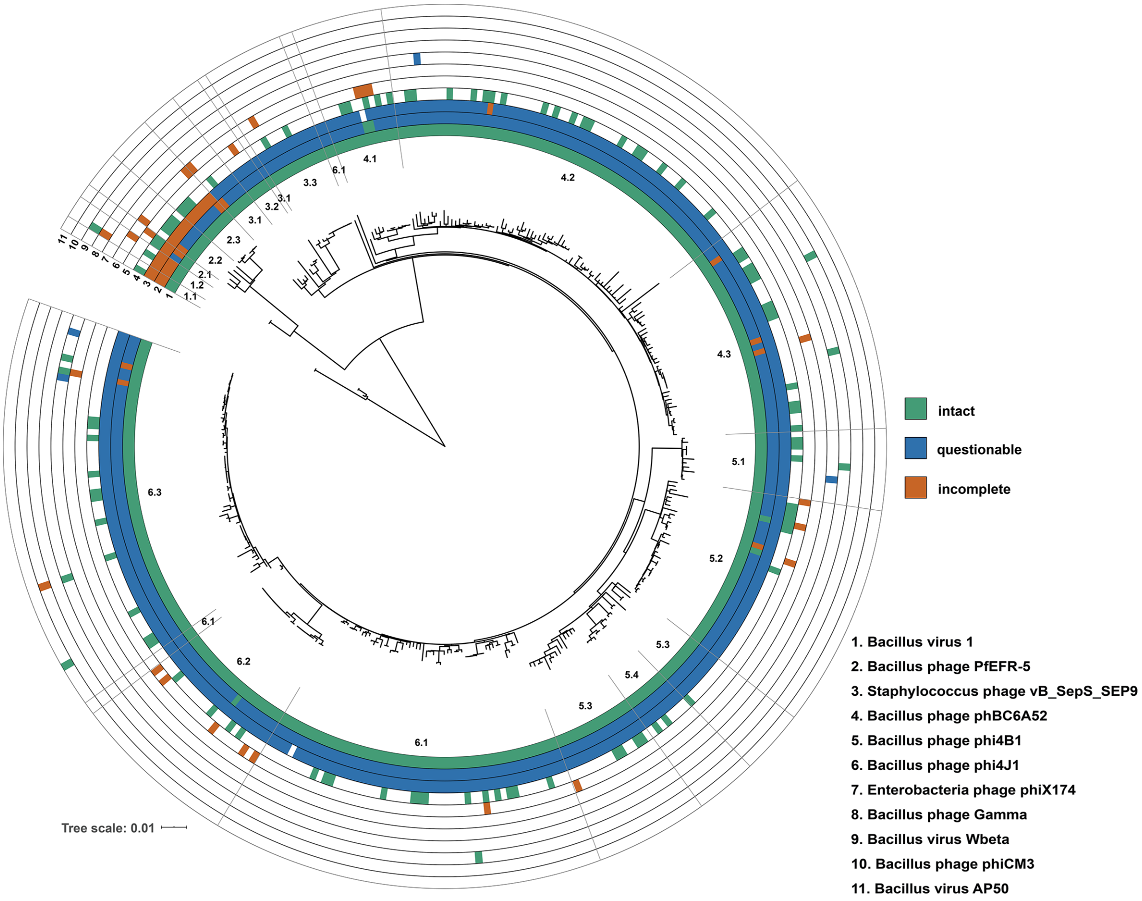

Fig. 2. Prophage sequences identified in whole-chromosome tree of 356 global *B. anthracis* isolates. Primary clusters are divided into their numbered nested clades by grey lines. The key on the right indicates the phage sequence present for each isolate, numbered according to their order from the inside of the ring to the outside. Colour indicates whether the phage sequence was determined to be intact, questionable or incomplete. Criteria related to this categorization can be found in Table S1. See Fig. S1 for a key to previously established classification schemes.

the confusion matrix was 16.89%. Applying the model to the validation dataset and comparing the observations and predictions, the overall accuracy was 0.861. The model always failed to predict primary cluster 1 (C Branch) for the validation dataset (*N*=1), and primary cluster 2 (B Branch) for both the training set (*N*=9) and the validation dataset (*N*=2), probably due to the reduced number of representatives comprising these clusters. The AMR gene *vml*R, the isolation source (host, environment or industry) and the continent of isolation were the most important variables in explaining genomic clusters across the entire dataset (Fig. 3a). The variable importance based on the mean decrease in accuracy for each individual cluster is shown in Fig. 3b. The absence of AMR gene *fos*B was the strongest predictor for primary cluster 1 (C Branch), whereas the *vml*R gene in primary cluster 2 (B Branch) acted as the strongest predictor. Nevertheless, both of these models exhibited negligible accuracy in the confusion matrix, suggesting more data are needed for accurate classification for these two clusters (Table S6). In cluster 3 (Ancient A) the continent of isolation was the strongest predictor by a large margin, as the vast majority of the samples that make up this population were isolated in Africa. For cluster 4 (Vollum) isolation source was the strongest predictor, followed by continent, as the majority of the isolates from this cluster were collected from industry (textile factories, animal processing plants, etc.) in North America. For cluster 5 (V770, Ames, Sterne, Aust94) the absence of the *vml*R gene was the strongest, lone overall predictor. Finally, in cluster 6 (TEA), isolation source, presence of the *vml*R gene and continent all showed comparatively strong power in classifying this cluster, with the majority of isolates from this cluster being isolated from animal hosts in North America and Europe.

Role of selection in shaping the *B. anthracis* genome

The program Gubbins predicted two instances of recombination, the first in a single isolate from Thailand [clade 5.2 (Ames), NCBI Sequence Read Archive: SRR5811219], based on 26 SNPs, and the second encompassing 13 isolates [comprising all of primary cluster 2 (B Branch)], based on

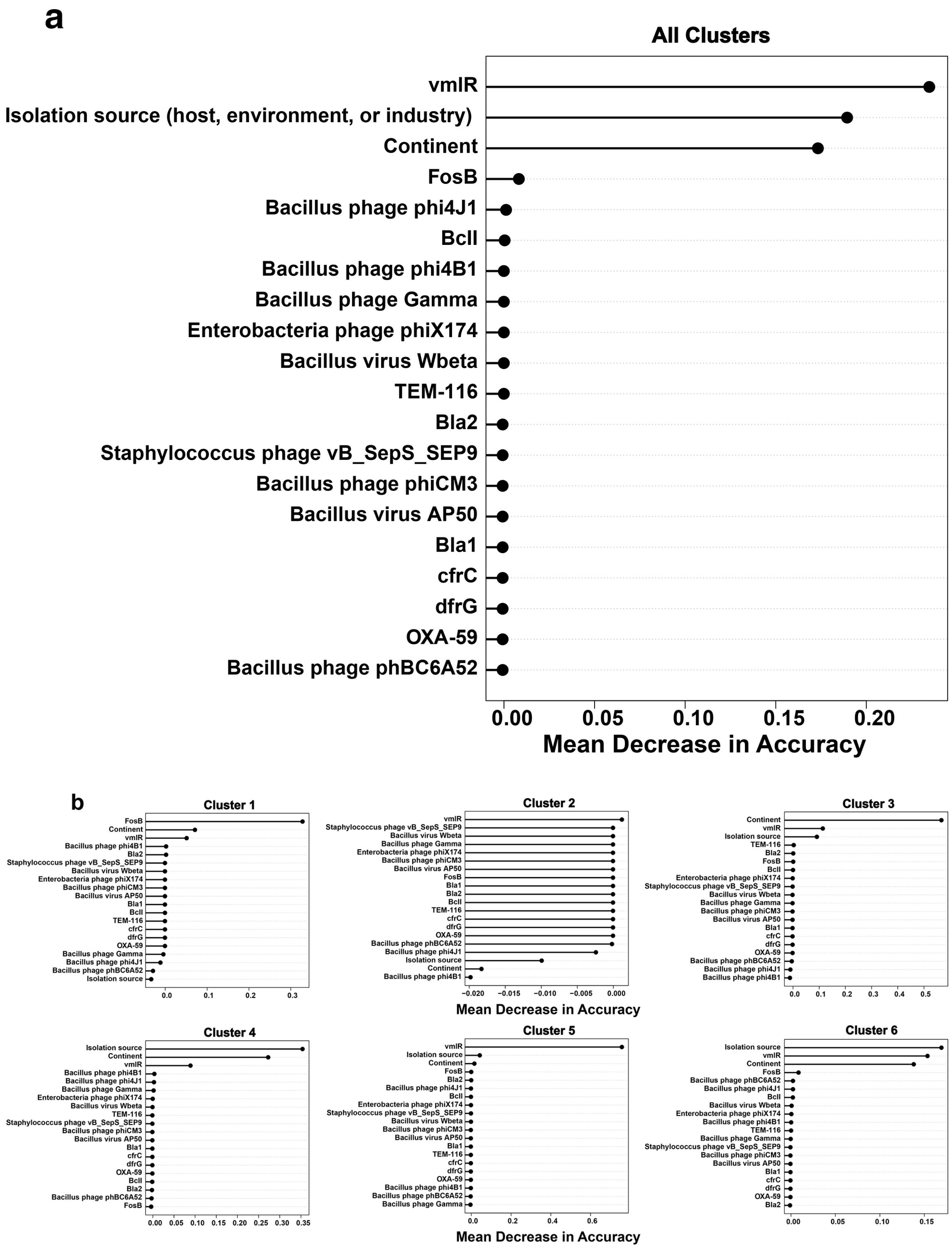

Fig. 3. Importance of the covariates in defining population genomic architecture for all primary clusters combined (a) and for each primary cluster on its own (b) by the RF classifier.

Table 2. Information for SNPs that exhibit a potentially high-impact effect and fall within clustered mutations across the *B. anthracis* genome; positions are relative to the Ames Ancestor reference genome (NCBI accession: AE017334.2)

Cluster	Position	Reference	Alternate	Comparative effect	Gene/product
1	1260604	C	T	Stop gained	DNA-binding response regulator
	1292469	C	A	Stop gained	Stage 0 sporulation regulatory protein
2	3140849	A	T	Stop gained	TPR domain protein
5 and 6	1748642	A	T	Stop gained	Chlorohydrolase family protein
	2423864	T	C	Start lost	Hypothetical protein

24 SNPs. Both instances of predicted recombination were specific to the *rrsA* rRNA gene [positions 9335–10841 in the Ames Ancestor reference genome (NCBI accession: AE017334.2)], which encodes the 16S rRNA, essential to translating messenger RNA into proteins.

After removing SNPs associated with recombination, SNP calls were split by primary cluster designations using a hierarchical approach, grouping primary clusters based on their nested structure, while filtering at a minimum allele frequency of 0.10 to avoid the identification of relatively rare alleles that were not necessarily indicative of their respective population genomic cluster. High-impact SNPs were then identified using the program snpEff. A detailed account of all 62 high-impact SNPs identified (including their predicted effect) is presented in Table S7. We also looked at clustered mutations in the same hierarchical manner, leading to the identification of 122 candidate genes potentially influencing selection (Table S8). Comparing both methodologies, five genes that spanned clustered mutations also contained a high-impact SNP (Table 2). These include genes coding for a DNA-binding response regulator and stage 0 sporulation regulatory protein (both specific to primary cluster 1), a tetratricopeptide repeat (TPR) domain protein [specific to primary cluster 2 (B Branch)], as well as a chlorohydrolase family protein and a hypothetical protein [both specific to primary clusters 5 (V770, Ames, Sterne, Aust94) and 6 (TEA)].

Lastly, we looked at non-synonymous mutations across the *B. anthracis* virulence genes (in the pXO1 and pXO2 plasmids), again using a minimum allele frequency of 0.10 to avoid the identification of relatively rare alleles that were not necessarily indicative of a group. All of the non-synonymous SNPs identified were on the pXO1 plasmid and spanned two toxin genes: the *cya* (calmodulin-sensitive adenylate cyclase) and the *pag*A (protective antigen) genes (Table 3). Both of the mutations in the *cya* gene were specific to clade 6.3 (WNA/TEABr011) for which all isolates were collected in western North America. In the *pag*A gene, one missense mutation was specific to the genetically and geographically diverse primary cluster 4 (Vollum), and the other to cluster 5 (V770, Ames, Sterne, Aust94), for which most of the isolates were collected in Asia and Europe.

DISCUSSION

Understanding the drivers of population genomic structure in pathogens is essential for making informed decisions related to wildlife management, disease control and public health. The data presented in this study offer the first detailed, global accounting of AMR genes and phage diversity in *B. anthracis*. In addition, our findings suggest that the six primary clusters defining population genomic structure in this species are consistent with differences in both AMR genes, geography and the source from which they were isolated. We also demonstrate that a recombination event linked to protein translation may take part in determining the persistence of certain *B. anthracis* strains. Finally, we offer a wealth of information on genomic diversity potentially associated with functional differences driving selection, allowing for further investigations into *B. anthracis* persistence, biogeography and evolution.

AMR has gained increased attention as a major threat to public health throughout the world [48, 49]. By documenting AMR genes on a global scale, we can gain a better understanding of how biogeography and persistence are transforming the genomic constitution of dangerous pathogens at both regional and wider scales [50, 51]. Based on our analysis of over 350 whole genomes, we have identified ten AMR genes present in *B. anthracis* isolates collected from over 35 countries, many

Table 3. Information for non-synonymous SNPs in virulence genes across the *B. anthracis* plasmids; positions are relative to the Ames Ancestor reference genome (NCBI accession: AE017336)

Cluster.clade	Position	Plasmid	Reference	Alternate	Comparative effect	Gene/product
6.3	123936	pXO1	A	T	Missense mutation	*cya*: calmodulin-sensitive adenylate cyclase
6.3	124007	pXO1	A	G	Missense mutation	*cya*: calmodulin-sensitive adenylate cyclase
4	145471	pXO1	C	T	Missense mutation	*pag*A: protective antigen
5	145577	pXO1	C	T	Missense mutation	*pag*A: protective antigen

consistent with the different population clusters examined in this study. Five of these genes are commonplace and can be found in the majority of isolates examined, whereas the other five are comparatively rare across the dataset. Given the application of commonly used antibiotic drugs, such as penicillin, doxycycline and ciprofloxacin, to treat *B. anthracis* infections, the regions where rare antibiotic-resistant gene isolates were sampled may benefit from monitoring, in order to document the persistence of these novel, resistant population clusters and modify antibiotic treatments for effectiveness [52, 53]. The resistance gene *bcII* for example, which was found in only six samples, is known to hydrolyse a large number of penicillins (Table 1). Rarer antibiotic-resistant gene strains such as these may be indicative of a larger problem with antibiotic resistance in other dangerous pathogens as well, especially if the overuse of certain antibiotics is driving resistance in those regions where novel resistance genes reside [54].

The influence of bacteriophage sequences on population genomic structure across the global dataset is less clear. As with AMR genes, several phage sequences were commonplace across isolates examined, while others were rarer or without pattern. Phage diversity was the least important factor in predicting population genomic structure based on the RF technique applied in this study. This is in contrast to studies of other pathogens, where phage sequence variation has been consistent with population genomic structure and therefore used for strain typing [55, 56]. Although previous studies have suggested some phage sequences may affect certain bacterial processes in *B. anthracis*, such as sporulation [17, 18], there was not an observable example of this leading to any advantage reflected in the form of genetically similar population clusters.

Applying the RF model, population genomic structure was most readily described by a combination of AMR genes, isolation location and source. The strongest predictor of population genomic structure when examining the dataset in its entirety was the presence of the AMR gene *vml*R, which was completely absent in primary cluster 5 [which was A.Br.001–A.Br.004 (Ames, Sterne, Aust94, V770) in the original classification system], the most genetically diverse population cluster examined in this study from which isolates were collected across Europe, Asia, Africa and the Americas. Interestingly, isolation source (host, environment or industry) was the second strongest predictor, suggesting that some strains of *B. anthracis* may be better suited to different environmental circumstances (or at least more readily cultured within them). Previous work that has examined population genomic structure has suggested that environmental growth outside of the host is possible [28]. Additionally, strains collected from industry may represent geographical consistencies in raw wool procural rather than a niche associated with this type of artificial environment [57, 58]. Nevertheless, long latent periods in the spore phase may be hindering our ability to detect environmental consistencies with population genomic structure. Not surprisingly, the continent of isolation was also a strong predictor in terms of population genomic structure, consistent with expected biogeographical patterns based on centuries of dispersal, complex trading patterns and global commerce. These findings are largely consistent with past work that has examined the population genetics and ubiquitous dissemination of this bacterium [26, 28, 58]. These combined forces – AMR genes, isolation source and biogeography – all seem to play a role in defining modern population structure in this bacterium.

Using RF models to look at the factors influencing each primary cluster individually, we found that varying circumstances seem to act as predictors for each individual cluster. The most underrepresented group, primary cluster 1, previously referred to as the C Branch in the *B. anthracis* literature and viewed as a rarely occurring clade [26, 28], is largely defined by the absence of the AMR gene *fos*B, which is found universally across all other population clusters examined. The relatively rare primary cluster 2 (B Branch) was not easily defined by any of the variables examined. Nevertheless, classification performance for both primary cluster 1 and cluster 2 was equally poor when assessing the accuracy. Previous work that has specifically examined isolates belonging to cluster 2 from Kruger National Park found that they were prevalent in more alkaline calcium-rich soils than cluster 3 (Ancient A) isolates occurring in the same region [30]. Cluster 3 (Ancient A) was described primarily by its isolation from the continent of Africa (although there are several isolates from elsewhere as well), suggesting that isolates from this group may be uniquely suited to or may have originated in this region. Primary cluster 4 is primarily described by a combination of isolation source and continent. This group, formerly referred to as A.Br.007 or Vollum in the literature, was isolated almost exclusively in a manufacturing setting in North America. Metadata and historical records for some of these isolates which were originally sequenced by the Centers for Disease Control (CDC) suggest that these isolates may have originated in other areas, most notably Asia and the Middle East [58, 59]. Cluster 5 (A.Br 001–004) is most readily described by the complete absence of the AMR gene *vml*R. Lastly, cluster 6 [previously the A.Br.008 and A.Br.009 lineages (TAE)] was primarily described by isolation source, as the majority of these isolates were collected from animal hosts throughout Europe and North America, although this group also contained isolates from Asia and South America in smaller numbers.

When examining population genomic structure in the context of candidate genes for selection, we see that recombination specific to primary cluster 2 (previously known as B Branch) may be responsible for the comparatively extreme difference in population structure in this group when compared to groups 3 to 6 (A Branch). A study that specifically looked at this group suggests that there may be phenotypic differences leading to contrasting mechanisms of infection, making this group specifically well suited to bovine species [58]. Given that this recombination event is rooted in a gene responsible for protein translation, these results support the hypothesis that phenotypic and functional traits for this cluster may be substantially different from the others.

We examined genes that were identified using two methods for pinpointing candidates for selection (high-impact SNPs and clustered mutations) and found that a range of functional differences may be driving population genomic structure. Primary cluster 1 (C Branch) exhibited premature stop codons in two genes, a DNA-binding response regulator and a stage 0 sporulation regulatory protein. If these premature stop codons are hindering this cluster's ability to produce proteins and influencing the timing and magnitude of sporulation, then this may indeed be why they are so underrepresented in the global dataset, and comparatively rare. Primary cluster 2 (B Branch) exhibited a premature stop codon in the TPR domain protein. TPR proteins may act as scaffolds for the assembly of different multiprotein complexes [60]. A premature stop codon in this sequence may be similarly affecting primary cluster 2's ability to persist and reproduce, leading to its similar rarity across the remainder of the global dataset (*N*=13/356). When primary clusters 5 and 6 are examined as a unit we see that the chlorohydrolase family protein exhibits a premature stop codon. Hydrolase proteins commonly perform as biochemical catalysts that use water to break a chemical bond, which typically results in dividing larger molecules into smaller molecules [61]. If this protein lacks the ability to perform this function, isolates specific to this group may be functionally different from the other population groupings. Overall these findings lay the groundwork for future studies into *B. anthracis* evolution, allowing for investigations into how protein structure drives functional and phenotypic differences across varied lineages.

Lastly, we looked at the *B. anthracis* virulence genes and found that several missense mutations may be influencing protein structure in some population clusters relative to others. Primary clusters 4 (Vollum) and 5 (A.Br001-004), the second and third most common designations across all isolates examined, exhibited different missense mutations in the *pag*A gene. The *pag*A gene encodes the protective antigen (PA), which binds to a receptor in sensitive eukaryotic cells, thereby facilitating the translocation of the enzymatic toxin components, oedema factor and lethal factor, across the target cell membrane [62]. Past work on this gene found six different haplotypes, which translate into three different amino acid sequences. Amino acid changes were shown to be located in an area near a highly antigenic region critical to lethal factor binding [63]. These mutations may therefore explain these clusters' comparatively robust prevalence compared to some others if this differentiated structure is more beneficial to genotypic persistence. We also found two mutations in the *cya* gene specific to clade 6.3 (WNA) entirely from North America. The *cya* gene codes for the calmodulin-sensitive adenylate cyclase that, when associated with PA, causes oedema. This protein product is not toxic in and of itself, although it is required for the survival of germinated spores within macrophages at the early stages of infection, provoking dramatic elevation of intracellular cAMP levels in the host [64].

When evaluating the population genomic structure of *B. anthracis* in light of biogeography, AMR, phage diversity and candidate genes for selection, we find varying explanations for differences in population genomic structure. Nevertheless, it should be noted that in a mined dataset such as this, inaccuracy in metadata and/or sequencing has the potential to produce unintentional errors. In addition, our dataset is highly biased towards developed countries where whole genome sequencing technology is readily available and government support for such work is more abundant. Given the complex dispersal history of this notorious pathogen and the competing factors that ultimately sculpt its global genomic architecture, no single factor alone can be attributed to its modern genomic constitution. Despite these limitations we were able to determine the most influential factors consistent with differences and similarities among lineages using modern bioinformatic techniques. The information provided in this study not only offers a detailed accounting of AMR genes and phage diversity in this species, but also allows for the groundwork upon which future *B. anthracis* studies into evolution can be built. This work has the potential to drive further discovery of functional differences in terms of virulence and genotypic persistence that may ultimately help to inform management strategies in the realm of public health and wildlife conservation.

Funding information
This work was supported by the NSF Division of Environmental Biology, grant number 1816161/2106221 (W.C.T.) and funds from the University at Albany (W.C.T.). P.L.K. was supported by the USDA National Institute of Food and Agriculture Hatch project number ME021908, through the Maine Agricultural and Forest Experiment Station.

Acknowledgements
We would like to thank Information Technology Services at the University at Albany, where all bioinformatic analyses were performed using the high-performance computing (HPC) cluster. Any use of trade, firm or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Author contributions

S. A. B., performed the research, analysed the data and wrote the manuscript. Y. H. H., provided additional analysis and improved the manuscript. P. L. K. and H. V. H., both provided major inputs to the manuscript structure, interpreted the results and improved the manuscript style. W. C. T., managed the project and developed the study, contributed to the research design, and improved the manuscript.

Conflicts of interest

The authors declared that there are no Conflicts of interest

References

1. Buckee CO, Jolley KA, Recker M, Penman B, Kriz P, *et al*. Role of selection in the emergence of lineages and the evolution of virulence in *Neisseria meningitidis*. *Proc Natl Acad Sci U S A* 2008;105:15082–15087.
2. Ernst CM, Braxton JR, Rodriguez-Osorio CA, Zagieboylo AP, Li L, *et al*. Adaptive evolution of virulence and persistence in carbapenem-resistant *Klebsiella pneumoniae*. *Nat Med* 2020;26:705–711.
3. Patel S. Drivers of bacterial genomes plasticity and roles they play in pathogen virulence, persistence and drug resistance. *Infection, Genetics and Evolution* 2016;45:151–164.
4. Gagneux S. Ecology and evolution of *Mycobacterium tuberculosis*. *Nat Rev Microbiol* 2018;16:202–213.
5. Myers JH, Cory JS. Ecology and evolution of pathogens in natural populations of Lepidoptera. *Evol Appl* 2016;9:231–247.

6. Brown SP, Cornforth DM, Mideo N. Evolution of virulence in opportunistic pathogens: generalism, plasticity, and control. *Trends Microbiol* 2012;20:336–342.
7. King KM. Pathogen population biology research can reduce international threats to tree health posed by invasive fungi. *outlook pest man* 2019;30:5–9.
8. Brockhurst MA, Buckling A, Rainey PB. The effect of a bacteriophage on diversification of the opportunistic bacterial pathogen, *Pseudomonas aeruginosa*. *Proc Biol Sci* 2005;272:1385–1391.
9. Kumar N, Lad G, Giuntini E, Kaye ME, Udomwong P, *et al*. Bacterial genospecies that are not ecologically coherent: population genomics of *Rhizobium leguminosarum*. *Open Biol* 2015;5:140133.
10. Collery MM, Kuehne SA, McBride SM, Kelly ML, Monot M, *et al*. What's a SNP between friends: The influence of single nucleotide polymorphisms on virulence and phenotypes of *Clostridium difficile* strain 630 and derivatives. *Virulence* 2017;8:767–781.
11. Liang X, Zhu J, Zhao Z, Zheng F, Zhang E, *et al*. A single nucleotide polymorphism is involved in regulation of growth and spore formation of Bacillus anthracis Pasteur II strain. *Front Cell Infect Microbiol* 2017;7:270.
12. Ekblom R, Galindo J. Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity (Edinb)* 2011;107:1–5.
13. Turner WC, Kamath PL, van Heerden H, Huang Y-H, Barandongo ZR, *et al*. The roles of environmental variation and parasite survival in virulence-transmission relationships. *R Soc Open Sci* 2021;8:210088.
14. Pfeilmeier S, Caly DL, Malone JG. Bacterial pathogenesis of plants: future challenges from a microbial perspective: challenges in bacterial molecular plant pathology. *Mol Plant Pathol* 2016;17:1298–1313.
15. Varela AR, Manaia CM. Human health implications of clinically relevant bacteria in wastewater habitats. *Environ Sci Pollut Res Int* 2013;20:3550–3569.
16. De Vos V, van Heerden H, Turner WC. Anthrax. Coetzer J, Thomson G, Maclachlan N and Penrith M-L (eds). In: *infectious diseases of livestock*, 3rd edn. Anipedia; 2018. http://www.anipedia.org/resources/anthrax/1203
17. Schuch R, Fischetti VA. Detailed genomic analysis of the Wβ and γ phages infecting *Bacillus anthracis*: implications for evolution of environmental fitness and antibiotic resistance. *J Bacteriol* 2006;188:3037–3051.
18. Schuch R, Fischetti VA. The secret life of the anthrax agent *Bacillus anthracis*: bacteriophage-mediated ecological adaptations. *PloS one* 2009;4:e6532.
19. Doĝanay M, Aydin N. Antimicrobial susceptibility of *Bacillus anthracis*. *Scand J Infect Dis* 1991;23:333–335.
20. White DG, Zhao S, Simjee S, Wagner DD, McDermott PF. Antimicrobial resistance of foodborne pathogens. *Microbes Infect* 2002;4:405–412.
21. Morgan TJ, Herman MA, Johnson LC, Olson BJ, Ungerer MC. Ecological Genomics: genes in ecology and ecology in genes. *Genome* 2018;61:v–vii.
22. Chen Y, Tenover FC, Koehler TM. β-Lactamase gene expression in a penicillin-resistant *Bacillus anthracis* strain. *Antimicrob Agents Chemother (Bethesda)* 2004;48:4873–4877.
23. Zhang E, Zhang H, He J, Li W, Wei J. Genetic diversity of *Bacillus anthracis* Ames lineage strains in China. *BMC Infect Dis* 2020;20:1–7.
24. Khmaladze E, Su W, Zghenti E, Buyuk F, Sahin M, *et al*. Molecular genotyping of *Bacillus anthracis* strains from Georgia and northeastern part of Turkey. *J Bacteriol Mycol* 2017;4.
25. Rondinone V, Serrecchia L, Parisi A, Fasanella A, Manzulli V, *et al*. Genetic characterization of *Bacillus anthracis* strains circulating in Italy from 1972 to 2018. *PloS one* 2020;15:e0227875.
26. Van Ert MN, Easterday WR, Simonson TS, U'Ren JM, Pearson T, *et al*. Strain-specific single-nucleotide polymorphism assays for the *Bacillus anthracis* Ames strain. *J Clin Microbiol* 2007;45:47–53.
27. Bruce SA, Schiraldi NJ, Kamath PL, Easterday WR, Turner WC. A classification framework for *Bacillus anthracis* defined by global genomic structure. *Evol Appl* 2020;13:935–944.
28. Sahl JW, Pearson T, Okinaka R, Schupp JM, Gillece JD, *et al*. A *Bacillus anthracis* genome sequence from the Sverdlovsk 1979 autopsy specimens. *MBio* 2016;7.
29. Pearson T, Busch JD, Ravel J, Read TD, Rhoton SD, *et al*. Phylogenetic discovery bias in *Bacillus anthracis* using single-nucleotide polymorphisms from whole-genome sequencing. *Proc Natl Acad Sci U S A* 2004;101:13536–13541.
30. Smith KL, DeVos V, Bryden H, Price LB, Hugh-Jones ME, *et al*. *Bacillus anthracis* diversity in kruger national park. *J Clin Microbiol* 2000;38:3780–3784.
31. Carlson CJ, Kracalik IT, Ross N, Alexander KA, Hugh-Jones ME, *et al*. The global distribution of Bacillus anthracis and associated anthrax risk to humans, livestock and wildlife. *Nat Microbiol* 2019;4:1337–1343.
32. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, *et al*. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19:455–477.
33. McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, *et al*. The comprehensive antibiotic resistance database. *Antimicrob Agents Chemother* 2013;57:3348–3357.
34. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, *et al*. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* 2016;44:W16-21.
35. Letunic I, Bork P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* 2007;23:127–128.
36. Adobe Inc. Adobe Illustrator. 2019. https://adobe.com/products/illustrator
37. Liaw A, Wiener M. Classification and regression by randomForest. *R news* 2002;2:18–22.
38. Barrett T, Clark K, Gevorgyan R, Gorelenkov V, Gribov E, *et al*. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res* 2012;40:D57-63.
39. Chuang LC, Kuo PH. Building a genetic risk model for bipolar disorder from genome-wide association data with random forest algorithm. *Sci Rep* 2017;7:1–0.
40. Lind AP, Anderson PC. Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties. *PloS one* 2019;14:e0219774.
41. Breiman L. *Manual on Setting Up, Using, and Understanding Random Forests V3. 1*. Berkeley, CA, USA: Statistics Department University of California; 2002, p. 58.
42. Li J. A new R package for spatial predictive modelling: SPM. Proceedings of the user 2018.
43. Svetnik V, Liaw A, Tong C, Wang T. Application of breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. In: *International workshop on multiple Classifier systems*. Berlin, Heidelberg: Springer, 2004. pp. 334–343.
44. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, *et al*. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* 2015;43:e15.
45. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, *et al*. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* 2012;6:80–92.
46. Cui Y, Schmid BV, Cao H, Dai X, Du Z, *et al*. Evolutionary selection of biofilm-mediated extended phenotypes in Yersinia pestis in response to a fluctuating environment. *Nat Commun* 2020;11:1–8.
47. Zhou T, Enyeart PJ, Wilke CO. Detecting clusters of mutations. *PLoS One* 2008;3:e3765.
48. Lehtinen S, Blanquart F, Lipsitch M, Fraser C, with the Maela Pneumococcal Collaboration. On the evolutionary ecology of multidrug resistance in bacteria. *PLoS Pathog* 2019;15:e1007763.

49. Sekyere JO, Asante J. Emerging mechanisms of antimicrobial resistance in bacteria and fungi: advances in the era of genomics. *Future Microbiol* 2018;13:241–262.

50. Sandulescu O. Global distribution of antimicrobial resistance in *E. coli. J Contemp Clin Pract* 2016;2:69–75.

51. Agersø Y, Andersen VD, Helwigh B, Høg BB, Jensen LB, *et al*. *DANMAP 2012: Use of Antimicrobial Agents and Occurrence of Antimicrobial Resistance in Bacteria from Food Animals, Food and Humans in Denmark*.

52. Heine HS, Shadomy SV, Boyer AE, Chuvala L, Riggins R, *et al*. Evaluation of combination drug therapy for treatment of antibiotic-resistant inhalation anthrax in a murine model. *Antimicrob Agents Chemother (Bethesda)* 2017;61.

53. Kelly DJ, Chulay JD, Mikesell P, Friedlander AM. Serum concentrations of penicillin, doxycycline, and ciprofloxacin during prolonged therapy in rhesus monkeys. *J Infect Dis* 1992;166:1184–1187.

54. Mather AE, Reeve R, Mellor DJ, Matthews L, Reid-Smith RJ, *et al*. Detection of rare antimicrobial resistance profiles by active and passive surveillance approaches. *Plos one* 2016;11:e0158515.

55. Neufeld T, Schwartz-Mittelmann A, Biran D, Ron EZ, Rishpon J. Combined phage typing and amperometric detection of released enzymatic activity for the specific identification and quantification of bacteria. *Anal Chem* 2003;75:580–585.

56. Uelze L, Grützke J, Borowiak M, Hammerl JA, Juraschek K, *et al*. Typing methods based on whole genome sequencing data. *One Health Outlook* 2020;2:1–9.

57. Irenge LM, Gala JL. Rapid detection methods for *Bacillus anthracis* in environmental samples: a review. *Appl Microbiol Biotechnol* 2012;93:1411–1422.

58. Pilo P, Frey J. Pathogenicity, population genetics and dissemination of *Bacillus anthracis*. *Infect Genet Evol* 2018;64:115–125.

59. Derzelle S, Aguilar-Bultet L, Frey J. Comparative genomics of *Bacillus anthracis* from the wool industry highlights polymorphisms of lineage A.Br.Vollum. *Infect Genet Evol* 2016;46:50–58.

60. Whitfield C, Mainprize IL. TPR motifs: hallmarks of a new polysaccharide export scaffold. *Structure* 2010;18:151–153.

61. Quinn JP, Kulakova AN, Cooley NA, McGrath JW. New ways to break an old bond: the bacterial carbon–phosphorus hydrolases and their role in biogeochemical phosphorus cycling. *Environ Microbiol* 2007;9:2392–2400.

62. Koehler TM. *Bacillus anthracis* genetics and virulence gene regulation. *Anthrax* 2002:143–164.

63. Price LB, Hugh-Jones M, Jackson PJ, Keim P. Genetic diversity in the protective antigen gene of *Bacillus anthracis*. *J Bacteriol* 1999;181:2358–2362.

64. Pezard C, Berche P, Mock M. Contribution of individual toxin components to virulence of Bacillus anthracis. *Infect Immun* 1991;59:3472–3477.