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[1,2], Pauline L. Kamath[3], Henriette van Heerden[4], Wendy C. Turner[5,1]

[1]Department of Biological Sciences, University at Albany – State University of New York, Albany, New York 12222, USA

[2]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

*Corresponding author

E-mail addresses:

SAB: sbruce@albany.edu

YHH: yenhua.huang@wisc.edu

PLK: pauline.kamath@maine.edu

HVH: henriette.vanheerden@up.ac.za

WCT: wendy.turner@wisc.edu

Keywords: bacterial genomics, natural selection, *Bacillus anthracis*, AMR

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 eleven 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 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.

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 presented data allows 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 emphasizes the importance of examining multiple factors dictating pathogen evolution and genotypic persistence.

Data Summary

- All NCBI accession numbers related to sequence reads and bioproject data used in this study are listed in Supplemental File 1.
- 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 (Ernst et al., 2020; Patel, 2016; Buckee et al., 2008). Phenotypic and genetic variation in a population may influence ecological composition and function, leading to increased or decreased evolutionary capacity under altered habitat regimes (Gagneux, 2018; Myers & Cory, 2016). Incorporating intraspecific diversity into effective management strategies demands the identification of factors influencing ecological plasticity and reproductive success (King, 2019; Brown et al., 2012). A wide range of genomic analyses have revealed genetic anomalies supporting ecologically variable phenotypes, suggesting a

consequential role for genomic architecture in driving intraspecific heterogeneity (Kumar et al., 2015; Brockhurst et al., 2005). For example, single nucleotide polymorphisms (SNPs) may facilitate the evolution of new phenotypes through the formation of novel proteins in regions that code for spore formation in *B. anthracis* or virulence factors in *Clostridium difficile* (Collery et al., 2017; Liang et al., 2017). By evaluating the relationship between whole genome architecture and ecologically relevant sequence variation, we can gain an acute understanding of how biocomplexity drives genomic structure in pathogenic bacterium (Ekblom & Galindo, 2011). Nevertheless, the relationship between genomic variation and adaptive relevance remains largely unknown for the vast majority of pathogenic species (Pfeilmeier et al., 2016; Varela & Manaia, 2013).

Bacillus anthracis has been extensively studied given its ability to cause anthrax, a disease that can be fatal to wildlife, livestock, and humans (De Vos et al., 2018). 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 (Schuch & Fischetti, 2009; Schuch & Fischetti, 2006). 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 bacterium including *B. anthracis* (White et al., 2002; Doğanay & Aydin, 1991). 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

93 emergence and reemergence in animals and humans (Morgan et al., 2018). Nevertheless, it
94 should be noted that the detection of an AMR gene does not always translate to conferred
95 resistance (Chen et al., 2004).
96 Individual and regional genetic diversity that differentiates *B. anthracis* populations by
97 SNP architecture has been identified on a global scale (Zhang et al., 2020; Rondinone et al.,
98 2020; Khmaladze et al., 2017; Van Ert et al. 2007). Recent work has refined our understanding
99 of population genomic structure for this species (Bruce et al., 2020; Sahl et al., 2016). Work by
100 Sahl et al. sought to expand on the original *B. anthracis* classification system, and generated a
101 SNP database used to characterize the branching structure of isolates based on 193 genomes
102 (Sahl et al., 2016). More recent genomic analyses that comprises the largest global phylogeny of
103 *B. anthracis* to date (356 genomes) has redefined *B. anthracis* population genomic structure,
104 resulting in six primary clusters and eighteen nested clades. This new classification system uses
105 an intuitive, simplified naming system and allows for linkable, rapid classification (Bruce et al.,
106 2020). Two of the major genotype clusters, cluster 1 (C Branch) and cluster 2 (B Branch) are
107 vastly underrepresented in terms of prevalence and have been hypothesized to be less fit than the
108 majority of *B. anthracis* specimens isolated and sequenced (Van Ert et al., 2007; Pearson et al.,
109 2004; Smith et al., 2000). However, the link between genomic architecture, and the scarcity of
110 these genotypes remains largely unexamined. In addition, some of the individual clades
111 identified are geographically specific, whereas others seem to be widely distributed, raising
112 numerous questions about what factors are driving evolutionary success in this species (Bruce et
113 al., 2020; Sahl et al., 2016; Van Ert et al., 2007). Understanding the relationships among spatial
114 variation, population stability, and genomic architectural variation is particularly important for *B.*
115 *anthracis*, as it is a major threat to wildlife, livestock, and public health globally (Carlson et al.,

2019). In this study, we explore genomic variation and selection in a global whole-genome dataset of *B. anthracis* isolates spanning thirty-nine countries and six continents. In addition, we apply an ensemble machine learning method (random forest) to elucidate the ways in which isolation source, geography, phage diversity, and AMR genes may be shaping genomic diversity and genotypic persistence. Random forest (RF) operates by constructing decision trees on various sub-samples of the dataset, allowing for predictions regarding evolutionary potential at the population level.

Methods

Whole genome mapping and assembly

The population genomic dataset used in these analyses was previously developed and published in Bruce et al. 2020 consisting of 356 *B. anthracis* whole genomes collected from the NCBI sequence read archive (Bruce et al., 2020; Supplemental File 1). 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. 2020.

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 (Bankevich et al., 2012). 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) (McArthur et al., 2013). 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) (Arndt et al., 2016). 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. 2020 was then annotated using iTOL (Letunic & Bork, 2007) 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 was then plotted geographically to understand patterns of resistance on a global scale using Adobe Illustrator (Adobe Inc., 2019).

Classifying population genomic architecture using random forest

To understand how various factors may be influencing the genomic architecture of *B. anthracis* we used a random forest approach (Liaw & Wiener, 2002), incorporating AMR gene data, phage diversity data, and isolate metadata (continent of isolation and source) accessed through the NCBI biosample database (Barret et al., 2012). RF has gained increased attention over the past several decades given its ability to produce excellent classification results while also being

computationally inexpensive (Lind & Anderson, 2019; Chuang & Kuo, 2017). 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 (Breiman, 2002). RF was carried out using the R package randomForest (Liaw & Wiener, 2002) to construct a multitude of decision trees and determine the mean prediction of each individual tree pertaining to the 6 primary population clusters (Bruce et al., 2020). The R package SPM was then used to carry out a 5-fold cross validation (Li , 2018).

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, 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 (Svetnik et al., 2004). 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 random forest model. To assess the model fit of the random forest we subjected the model to the validation dataset and estimated the accuracy. In order 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 OOB error (training observations not

184 included in the bootstrap) from a dataset with the selected variable permuted and the OOB error
185 from the original dataset (Breiman, 2002).
186
187 **Recombination, high impact SNPs, and candidate genes for selection**
188 To determine how recombination may be influencing population genomic structure across our
189 dataset we first used the program Gubbins to iteratively identify loci containing elevated
190 densities of base substitutions in the SNP dataset (prior to removal of recombinant sequences)
191 (Croucher et al., 2015).
192 To analyze selection in non-recombining regions, we analyzed SNPs (post-removal of
193 recombinant sequences) using the program SnpEff (Cingolani et al., 2012). SnpEff annotates and
194 predicts the effects of genetic variants on genes and proteins (such as amino acid changes). To
195 assess "high impact" SNPs influencing population genomic structure, we compiled a list of SNPs
196 that produce significant changes to protein structure in the *B. anthracis* chromosome and
197 plasmids, specific to each primary cluster and groups of primary clusters, such as mutations that
198 result in the gain of a stop codon, the loss of a start codon, and splice region variants. We also
199 looked for clustered SNPs across each of the aforementioned groups to identify genes that were
200 possibly associated with selection using a modified version of the algorithm developed by Cui et
201 al. 2020 (Cui et al., 2020), classifying genes that showed 3 or more mutations within an 2000bp
202 range, as well as genes that showed 2 or more SNPs within a 50bp range. Clustered mutations
203 have a low probability of occurring under a neutral substitution model, in which variations are
204 assumed to be randomly distributed across the genome (Zhou et al., 2008). Examining the ratio
205 of nonsynonymous to synonymous SNPs at the gene level was problematic given the clonal
206 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 analyzed (Fig. 1, Table 1). Additional information regarding the AMR genes and their frequency are provided in Table S3. A key linking the classification framework shown here to the previously established branch labels outlined by Sahl et al. 2016 are provided in Fig. S1. Five AMR genes (*mph*L, *bla*1, *fos*B, *bla*2, and *vml*R) 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 of the United States (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 United States ($N = 5$). *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 United States between 1956 and 1978 from clade 4.1 (Vollum, $N = 3$), as well as the entirety of 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 6 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 3 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 eleven 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 are 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 cluster 1 and 2 (C and B Branches; $N = 18$). The 8 remaining prophage sequences were scattered in comparatively minimal amounts across the global dataset,

253 with the exception of Bacillus phage phBC6A52 which was intact in a large number of the
254 isolates examined ($N = 91$), with seemingly no link to relatedness or geography among isolates.
255

256 ## Explaining global genomic clusters with random forest

257 The RF model was trained using 5-fold cross validation with a training dataset. The best model
258 parameter (where number of trees (Ntree) and number of variables (Mtry) equaled 400 and 9
259 respectively) produced a cross correlation rate that showed a high value of 83.6, while kappa
260 equaled 0.764. Variables examined include presence of AMR genes and phage sequences, as
261 well as sample source (details provided in Table S2). With this combination of Ntree and Mtry,
262 the out-of-bag (OOB) error based on the confusion matrix was 16.89%. Applying the model to
263 the validation dataset and comparing the observations and predictions, the overall accuracy was
264 0.861. The model always failed to predict primary cluster 1 (C Branch) for the validation dataset
265 ($N = 1$), and primary cluster 2 (B Branch) for both the training set ($N = 9$) and validation dataset
266 ($N = 2$), likely due to the reduced number of representatives comprising these clusters. The AMR
267 gene *vml*R, the isolation source (host, environment, or industry), and the continent of isolation
268 were the most important variables in explaining genomic clusters across the entire dataset (Fig.
269 3a). The variable importance based on the mean decrease in accuracy for each individual cluster
270 is shown in Fig. 3b. The absence of AMR gene *fos*B was the strongest predictor for primary
271 cluster 1 (C Branch), whereas the *vml*R gene in primary cluster 2 (B Branch) acted as the
272 strongest predictor. Nevertheless, both of these models exhibited negligible accuracy in the
273 confusion matrix, suggesting more data is needed for accurate classification for these two
274 clusters (Table S6). In cluster 3 (Ancient A) the continent of isolation was the strongest predictor
275 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 as the majority of isolates from this cluster were isolated from animal hosts in North America and Europe.

Selection's role 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 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 ribosomal RNA, 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 accounting of all 62 high impact SNPs identified (including their predicted effect) can be found 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) through 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 offers the first detailed, global accounting of AMR genes and phage diversity in this bacterium. In addition, our findings suggest that the 6 primary clusters defining population genomic structure in this species are consistent with differences in both AMR genes,

322 geography, and the source from which they were isolated. We also demonstrate that a
323 recombination event linked to protein translation may take part in determining the persistence of
324 certain *B. anthracis* strains. Finally, we offer a wealth of information on genomic diversity
325 potentially associated with functional differences driving selection, allowing for further
326 investigations into *B. anthracis* persistence, biogeography and evolution.
327 AMR has gained increased attention as a major threat to public health throughout the
328 world (Lehtinen et al., 2019; Sekyere & Asante 2018). By documenting AMR genes on a global
329 scale, we can gain a better understanding of how biogeography and persistence is transforming
330 the genomic constitution of dangerous pathogens at both regional and wider scales (Sandulescu,
331 2016; Agersø et al, 2013). Based on our analysis of over 350 whole genomes we have identified
332 ten AMR genes present in *B. anthracis* isolates collected from over 35 countries, many
333 consistent with the different population clusters examined in this study. Five of these genes are
334 commonplace and can be found in the majority of isolates examined, whereas the other five are
335 comparatively rare across the dataset. Given the application of commonly used antibiotic drugs,
336 such as penicillin, doxycycline, and ciprofloxacin to treat *B. anthracis* infections, the regions
337 where rare antibiotic resistant gene isolates were sampled may benefit from monitoring, in order
338 to document the persistence of these novel, resistant population clusters and modify antibiotic
339 treatments for effectiveness (Heine et al., 2017; Kelly et al., 1992). The resistance gene *bcII* for
340 example, which was found in only six samples is known to hydrolyze a large number of
341 penicillins (Table 1). Rarer antibiotic resistant gene strains such as these may be indicative of a
342 larger problem with antibiotic resistance in other dangerous pathogens as well, especially if the
343 overuse of certain antibiotics is driving resistance in those regions where novel resistance genes
344 reside (Mather et al., 2016).

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 random forest 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 (Uelze et al., 2020; Neufeld et al., 2003). Although previous studies have suggested some phage sequences may affect certain bacterial processes in *B. anthracis*, such as sporulation (Schuch & Fischetti, 2009; Schuch & Fischetti, 2006), there was not an observable example of this leading to any advantage reflected in the form of genetically similar population clusters.

Applying the random forest 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 (Sahl et al., 2016). Additionally, strains collected from industry may represent geographic consistencies in raw wool procural rather than a niche associated with this type of artificial environment (Pilo & Frey 2018; Irenge & Gala,

368 2012). Nevertheless, long latent periods in the spore phase may be hindering our ability to detect
369 environmental consistencies with population genomic structure. Not surprisingly the continent of
370 isolation was also a strong predictor in terms of population genomic structure, consistent with
371 expected biogeographic patterns based on centuries of dispersal, complex trading patterns and
372 global commerce. These findings are largely consistent with past work that has examined the
373 population genetics and ubiquitous dissemination of this bacterium (Pilo & Frey 2018; Sahl et
374 al., 2016; Van Ert et al., 2007). These combined forces—AMR genes, isolation source and
375 biogeography—all seem to play a role in defining modern population structure in this bacterium.
376 Using random forest models to look at the factors influencing each primary cluster
377 individually, we found that varying circumstances seem to act as predictors for each individual
378 cluster. The most underrepresented group, primary cluster 1, previously referred to as the C
379 Branch in the *B. anthracis* literature and viewed as a rarely occurring clade (Sahl et al., 2016;
380 Van Ert et al., 2007), is largely defined by the absence of the AMR gene *fos*B, which is found
381 universally across all other population clusters examined. The relatively rare primary cluster 2 (B
382 Branch) was not easily defined by any of the variables examined. Nevertheless, classification
383 performance for both primary cluster 1 and 2 were equally poor when assessing the accuracy.
384 Previous work that has specifically examined isolates belonging to cluster 2 from Kruger
385 National Park found that they were prevalent in more alkaline calcium rich soils than the cluster
386 3 (Ancient A) isolates occurring in the same region (Smith et al., 2000). Cluster 3 (Ancient A)
387 was described primarily by its isolation from the continent of Africa (although there are several
388 isolates from elsewhere as well) suggesting that isolates from this group may be uniquely suited
389 to or may have originated in this region. Primary cluster 4 is primarily described by a
390 combination of isolation source and continent. This group, formerly referred to as A.Br.007 or

391 Vollum in the literature, was isolated almost exclusively in a manufacturing setting in North
392 America. Metadata and historical records for some of these isolates which were originally
393 sequenced by the Center of Disease Control (CDC) suggest that these isolates may have
394 originated in other areas, most notably Asia and the Middle East (Pilo & Frey 2018; Derzelle et
395 al., 2016). Cluster 5 (A.Br 001-004) is most readily described by the complete absence of the
396 AMR gene *vml*R. Lastly, cluster 6 (previously the A.Br.008 and A.Br.009 lineages (TAE)) was
397 primarily described by isolation source, as the majority of these isolates were collected from
398 animal hosts throughout Europe and North America, although this group also contained isolates
399 from Asia and South America in smaller numbers.
400 When examining population genomic structure in the context of candidate genes for
401 selection, we see that recombination specific to primary cluster 2 (previously known as B
402 Branch) may be responsible for the comparatively extreme difference in population structure in
403 this group when compared to groups 3 through 6 (A Branch). A study that specifically looked at
404 this group suggests that there may be phenotypic differences leading to contrasting mechanisms
405 of infection, make this group specifically well suited to bovine species (Pilo & Frey 2018).
406 Given that this recombination event is rooted in a gene responsible for protein translation, these
407 results support the hypothesis that phenotypic and functional traits for this cluster may be
408 substantially different from the others.
409 We examined genes that were identified using two methods for pinpointing candidates
410 for selection (high-impact SNPs and clustered mutations) and found that a range of functional
411 differences may be driving population genomic structure. Primary cluster 1 (C Branch) exhibited
412 premature stop codons in two genes, a DNA-binding response regulator and a stage 0 sporulation
413 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 (Whitfield & Mainprize, 2010). 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 (Quinn et al., 2007). If this protein lacks the ability to perform this function, isolates specific to this group may be functionally different than 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 2nd and 3rd 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, edema factor and lethal factor, across the target cell membrane (Koehler, 2007). 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

437 factor binding (Price et al., 1999). These mutations may therefore explain these cluster's
438 comparatively robust prevalence compared to some others if this differentiated structure is more
439 beneficial to genotypic persistence. We also found two mutations in the *cya* gene specific to
440 clade 6.3 (WNA) entirely from North America. The *cya* gene codes for the calmodulin-sensitive
441 adenylate cyclase that, when associated with PA, causes edema. This protein product is not toxic
442 in and of itself, although it is required for the survival of germinated spores within macrophages
443 at the early stages of infection, provoking dramatic elevation of intracellular cAMP levels in the
444 host (Pezard et al., 1991).

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

459 virulence and genotypic persistence that may ultimately help to inform management strategies in
460 the realm of public health and wildlife conservation.
461

462 ## Author Statements

463 S.A.B. performed the research, analyzed the data and wrote the manuscript. Y.H.H. provided
464 additional analysis and improved the manuscript. P.L.K and H.V.H both provided major inputs
465 to the manuscript structure, interpreted the results and improved the manuscript style. W.C.T.
466 managed the project and developed the study, contributed to the research design, and improved
467 the manuscript.
468

469 ## Conflict of Interests

470 This manuscript contains no copyrighted material; nor have we have discussed its content with a
471 *Microbial Genomics* editor. None of the authors hold competing interests or affiliations with or
472 involvement in any organization or entity with any financial interest, or non-financial interest in
473 the subject matter or materials discussed in this manuscript.
474

475 ## Acknowledgements

476 We would like to thank Information Technology Services at the University at Albany, where all
477 bioinformatic analyses was performed using the high-performance computing (HPC) cluster.
478 Any use of trade, firm, or product names is for descriptive purposes only and does not imply
479 endorsement by the U.S. Government.
480

481 ## 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 & Forest Experiment Station.

References

Adobe Inc. Adobe Illustrator [Internet]. 2019. Available from: https://adobe.com/products/illustrator

Agersø Y, Andersen VD, Helwigh B, Høg BB, Jensen LB, Jensen VF, Korsgaard H, Larsen LS, Pedersen K, Seyfarth AM, Dalby T. DANMAP 2012: Use of antimicrobial agents and occurrence of antimicrobial resistance in bacteria from food animals, food and humans in Denmark.

Ågren J, Finn M, Bengtsson B, Segerman B. Microevolution during an anthrax outbreak leading to clonal heterogeneity and penicillin resistance. PLoS One. 2014 Feb 13;9(2):e89112.

Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, Wishart DS. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic acids research. 2016 May 3;44(W1):W16-21.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology. 2012 May 1;19(5):455-77.

Barrett T, Clark K, Gevorgyan R, Gorelenkov V, Gribov E, Karsch-Mizrachi I, Kimelman M, Pruitt KD, Resenchuk S, Tatusova T, Yaschenko E. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. Nucleic acids research. 2012 Jan 1;40(D1):D57-63.

Breiman L. Manual on setting up, using, and understanding random forests v3. 1. Statistics Department University of California Berkeley, CA, USA, 1, 58; 2002.

Brockhurst MA, Buckling A, Rainey PB. The effect of a bacteriophage on diversification of the opportunistic bacterial pathogen, Pseudomonas aeruginosa. Proceedings of the Royal Society B: Biological Sciences. 2005 Jul 7;272(1570):1385-91.

Brown SP, Cornforth DM, Mideo N. Evolution of virulence in opportunistic pathogens: generalism, plasticity, and control. Trends in microbiology. 2012 Jul 1;20(7):336-42.

Bruce SA, Schiraldi NJ, Kamath PL, Easterday WR, Turner WC. A classification framework for Bacillus anthracis defined by global genomic structure. Evolutionary Applications. 2020 May;13(5):935-44.

Buckee CO, Jolley KA, Recker M, Penman B, Kriz P, Gupta S, Maiden MC. Role of selection in the emergence of lineages and the evolution of virulence in Neisseria meningitidis. Proceedings of the National Academy of Sciences. 2008 Sep 30;105(39):15082-7.

Carlson CJ, Kracalik IT, Ross N, Alexander KA, Hugh-Jones ME, Fegan M, Elkin BT, Epp T, Shury TK, Zhang W, Bagirova M. The global distribution of Bacillus anthracis and associated anthrax risk to humans, livestock and wildlife. Nature microbiology. 2019 Aug;4(8):1337-43.

Chen Y, Tenover FC, Koehler TM. β-Lactamase gene expression in a penicillin-resistant Bacillus anthracis strain. Antimicrobial agents and chemotherapy. 2004 Dec 1;48(12):4873-7.

Chuang LC, Kuo PH. Building a genetic risk model for bipolar disorder from genome-wide association data with random forest algorithm. Scientific reports. 2017 Jan 3;7(1):1-0.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 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. 2012 Apr 1;6(2):80-92.

Collery MM, Kuehne SA, McBride SM, Kelly ML, Monot M, Cockayne A, Dupuy B, Minton NP. 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 Aug 18;8(6):767-81.

Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic acids research. 2015 Feb 18;43(3):e15-.

Cui Y, Schmid BV, Cao H, Dai X, Du Z, Easterday WR, Fang H, Guo C, Huang S, Liu W, Qi Z. Evolutionary selection of biofilm-mediated extended phenotypes in Yersinia pestis in response to a fluctuating environment. Nature communications. 2020 Jan 15;11(1):1-8.

De Vos V, van Heerden H, Turner WC. Anthrax. In: Infectious Diseases of Livestock. 3rd edition. Coetzer J.A.W., Thomson, G.R., Maclachlan, N.J. and M.-L. Penrith (editors); 2018. Anipedia, http://www.anipedia.org/resources/anthrax/1203.

Derzelle S, Aguilar-Bultet L, Frey J. Comparative genomics of Bacillus anthracis from the wool industry highlights polymorphisms of lineage A. Br. Vollum. Infection, genetics and evolution. 2016 Dec 1;46:50-8.

Doĝanay M, Aydin N. Antimicrobial susceptibility of Bacillus anthracis. Scandinavian journal of infectious diseases. 1991 Jan 1;23(3):333-5.

Ekblom R, Galindo J. Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity. 2011 Jul;107(1):1-5.

Ernst CM, Braxton JR, Rodriguez-Osorio CA, Zagieboylo AP, Li L, Pironti A, Manson AL, Nair AV, Benson M, Cummins K, Clatworthy AE. Adaptive evolution of virulence and persistence in carbapenem-resistant Klebsiella pneumoniae. Nature medicine. 2020 May;26(5):705-11.

Gagneux S. Ecology and evolution of Mycobacterium tuberculosis. Nature Reviews Microbiology. 2018 Apr;16(4):202.

Heine HS, Shadomy SV, Boyer AE, Chuvala L, Riggins R, Kesterson A, Myrick J, Craig J, Candela MG, Barr JR, Hendricks K. Evaluation of combination drug therapy for treatment of antibiotic-resistant inhalation anthrax in a murine model. Antimicrobial agents and chemotherapy. 2017 Sep 1;61(9).

Irenge LM, Gala JL. Rapid detection methods for Bacillus anthracis in environmental samples: a review. Applied microbiology and biotechnology. 2012 Feb;93(4):1411-22.

Kelly DJ, Chulay JD, Mikesell P, Friedlander AM. Serum concentrations of penicillin, doxycycline, and ciprofloxacin during prolonged therapy in rhesus monkeys. Journal of infectious Diseases. 1992 Nov 1;166(5):1184-7.

Khmaladze E, Su W, Zghenti E, Buyuk F, Sahin M, Nicolich MP, Baillie L, Obiso R, Kotorashvili A. Molecular genotyping of Bacillus anthracis strains from Georgia and northeastern part of Turkey. Journal of Bacteriology and Mycology. 2017 Sep 6;4(3).

King KM. Pathogen Population Biology Research can Reduce International Threats to Tree Health Posed by Invasive Fungi. Outlooks on Pest Management. 2019 Feb 1;30(1):5-9.

Koehler TM. Bacillus anthracis genetics and virulence gene regulation. Anthrax. 2002:143-64.

Kumar N, Lad G, Giuntini E, Kaye ME, Udomwong P, Shamsani NJ, Young JP, Bailly X. Bacterial genospecies that are not ecologically coherent: population genomics of Rhizobium leguminosarum. Open biology. 2015 Jan 1;5(1):140133.

Lehtinen S, Blanquart F, Lipsitch M, Fraser C, with the Maela Pneumococcal Collaboration. On the evolutionary ecology of multidrug resistance in bacteria. PLoS pathogens. 2019 May 13;15(5):e1007763.

Letunic I, Bork P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics. 2007 Jan 1;23(1):127-8.

Li J. A new R package for spatial predictive modelling: spm. Proceedings of the useR. 2018.

Liang X, Zhu J, Zhao Z, Zheng F, Zhang E, Wei J, Ji Y, Ji Y. A single nucleotide polymorphism is involved in regulation of growth and spore formation of Bacillus anthracis Pasteur II strain. Frontiers in cellular and infection microbiology. 2017 Jun 28;7:270.

Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002 Dec 3;2(3):18-22.

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 Jul 11;14(7):e0219774.

Mather AE, Reeve R, Mellor DJ, Matthews L, Reid-Smith RJ, Dutil L, Haydon DT, Reid SW. Detection of rare antimicrobial resistance profiles by active and passive surveillance approaches. Plos one. 2016 Jul 8;11(7):e0158515.

McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, Bhullar K, Canova MJ, De Pascale G, Ejim L, Kalan L. The comprehensive antibiotic resistance database. Antimicrobial agents and chemotherapy. 2013 Jul 1;57(7):3348-57.

Morgan TJ, Herman MA, Johnson LC, Olson BJ, Ungerer MC. Ecological Genomics: genes in ecology and ecology in genes. Genome. 2018;61(4):v-ii.

Myers JH, Cory JS. Ecology and evolution of pathogens in natural populations of Lepidoptera. Evolutionary Applications. 2016 Jan;9(1):231-47.

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. Analytical chemistry. 2003 Feb 1;75(3):580-5.

Patel S. Drivers of bacterial genomes plasticity and roles they play in pathogen virulence, persistence and drug resistance. Infection, Genetics and Evolution. 2016 Nov 1;45:151-64.

Pearson T, Busch JD, Ravel J, Read TD, Rhoton SD, U'ren JM, Simonson TS, Kachur SM, Leadem RR, Cardon ML, Van Ert MN. Phylogenetic discovery bias in Bacillus anthracis using single-nucleotide polymorphisms from whole-genome sequencing. Proceedings of the National Academy of Sciences. 2004 Sep 14;101(37):13536-41.

Pezard C, Berche P, Mock M. Contribution of individual toxin components to virulence of Bacillus anthracis. Infection and immunity. 1991 Oct 1;59(10):3472-7.

Pfeilmeier S, Caly DL, Malone JG. Bacterial pathogenesis of plants: future challenges from a microbial perspective: challenges in bacterial molecular plant pathology. Molecular plant pathology. 2016 Oct;17(8):1298-313.

Pilo P, Frey J. Pathogenicity, population genetics and dissemination of Bacillus anthracis. Infection, genetics and evolution. 2018 Oct 1;64:115-25.

Price LB, Hugh-Jones M, Jackson PJ, Keim P. Genetic Diversity in the Protective Antigen Gene ofBacillus anthracis. Journal of bacteriology. 1999 Apr 15;181(8):2358-62.

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. Environmental microbiology. 2007 Oct;9(10):2392-400.

Rondinone V, Serrecchia L, Parisi A, Fasanella A, Manzulli V, Cipolletta D, Galante D. Genetic characterization of Bacillus anthracis strains circulating in Italy from 1972 to 2018. PloS one. 2020 Jan 13;15(1):e0227875.

Sahl JW, Pearson T, Okinaka R, Schupp JM, Gillece JD, Heaton H, Birdsell D, Hepp C, Fofanov V, Noseda R, Fasanella A. A Bacillus anthracis genome sequence from the Sverdlovsk 1979 autopsy specimens. MBio. 2016 Nov 2;7(5).

Sandulescu O. Global distribution of antimicrobial resistance in E. coli. Journal of Contemporary Clinical Practice. 2016 Sep 1;2(2):69-75.

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. Journal of bacteriology. 2006 Apr 15;188(8):3037-51.

Schuch R, Fischetti VA. The secret life of the anthrax agent Bacillus anthracis: bacteriophage-mediated ecological adaptations. PloS one. 2009 Aug 12;4(8):e6532.

Sekyere JO, Asante J. Emerging mechanisms of antimicrobial resistance in bacteria and fungi: advances in the era of genomics. Future microbiology. 2018 Feb;13(2):241-62.

Smith KL, DeVos V, Bryden H, Price LB, Hugh-Jones ME, Keim P. Bacillus anthracis diversity in kruger national park. Journal of clinical microbiology. 2000 Oct 1;38(10):3780-4.

Svetnik V, Liaw A, Tong C, Wang T. Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. InInternational Workshop on Multiple Classifier Systems 2004 Jun 9 (pp. 334-343). Springer, Berlin, Heidelberg.

Uelze L, Grützke J, Borowiak M, Hammerl JA, Juraschek K, Deneke C, Tausch SH, Malorny B. Typing methods based on whole genome sequencing data. One Health Outlook. 2020 Dec;2(1):1-9.

Van Ert MN, Easterday WR, Simonson TS, U'Ren JM, Pearson T, Kenefic LJ, Busch JD, Huynh LY, Dukerich M, Trim CB, Beaudry J. Strain-specific single-nucleotide polymorphism assays for the Bacillus anthracis Ames strain. Journal of clinical microbiology. 2007 Jan 1;45(1):47-53.

Varela AR, Manaia CM. Human health implications of clinically relevant bacteria in wastewater habitats. Environmental Science and Pollution Research. 2013 Jun;20(6):3550-69.

White DG, Zhao S, Simjee S, Wagner DD, McDermott PF. Antimicrobial resistance of foodborne pathogens. Microbes and infection. 2002 Apr 1;4(4):405-12.

Whitfield C, Mainprize IL. TPR motifs: hallmarks of a new polysaccharide export scaffold. Structure. 2010 Feb 10;18(2):151-3.

Zhang E, Zhang H, He J, Li W, Wei J. Genetic diversity of Bacillus anthracis Ames lineage strains in China. BMC infectious diseases. 2020 Dec;20(1):1-7.

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 phosphotransferases that inactivates macrolides such as erythromycin, clarithromycin, azithromycin
*bla*1	antibiotic inactivation	ARO:3000090	A chromosomal-encoded beta-lactamase, which hydrolyzes penicillins
*fos*B	antibiotic inactivation	ARO:3000172	A thiol transferase that leads to the resistance of Fosfomycin
*bla*2	antibiotic inactivation	ARO:3004189	A chromosomal-encoded beta-lactamase, which has penicillin, cephalosporin, and carbapenem-hydrolizing 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 hydrolyzes 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

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	Ref	Alt	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 through 6	1748642	A	T	stop gained	chlorohydrolase family protein
	2423864	T	C	start lost	hypothetical protein

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	Ref	Alt	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

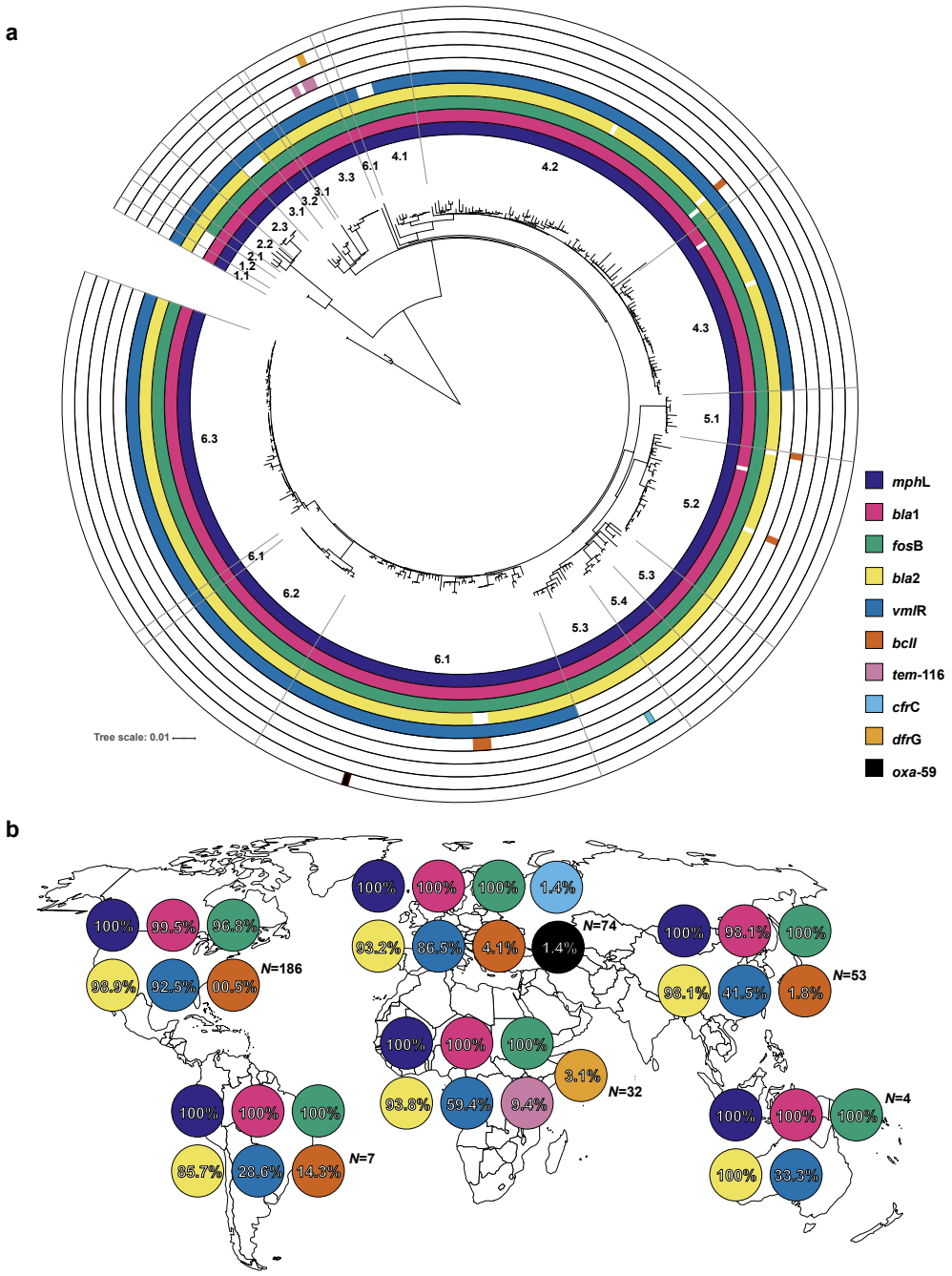

Figure 1. AMR genes identified in the whole-chromosome tree of 356 global *B. anthracis* isolates (a). Primary clusters are divided into their numbered nested clades by gray lines. The key on the right lists the 10 AMR genes identified. Outer rings reflect presence (color) or absence (white) of each 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 colored 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.

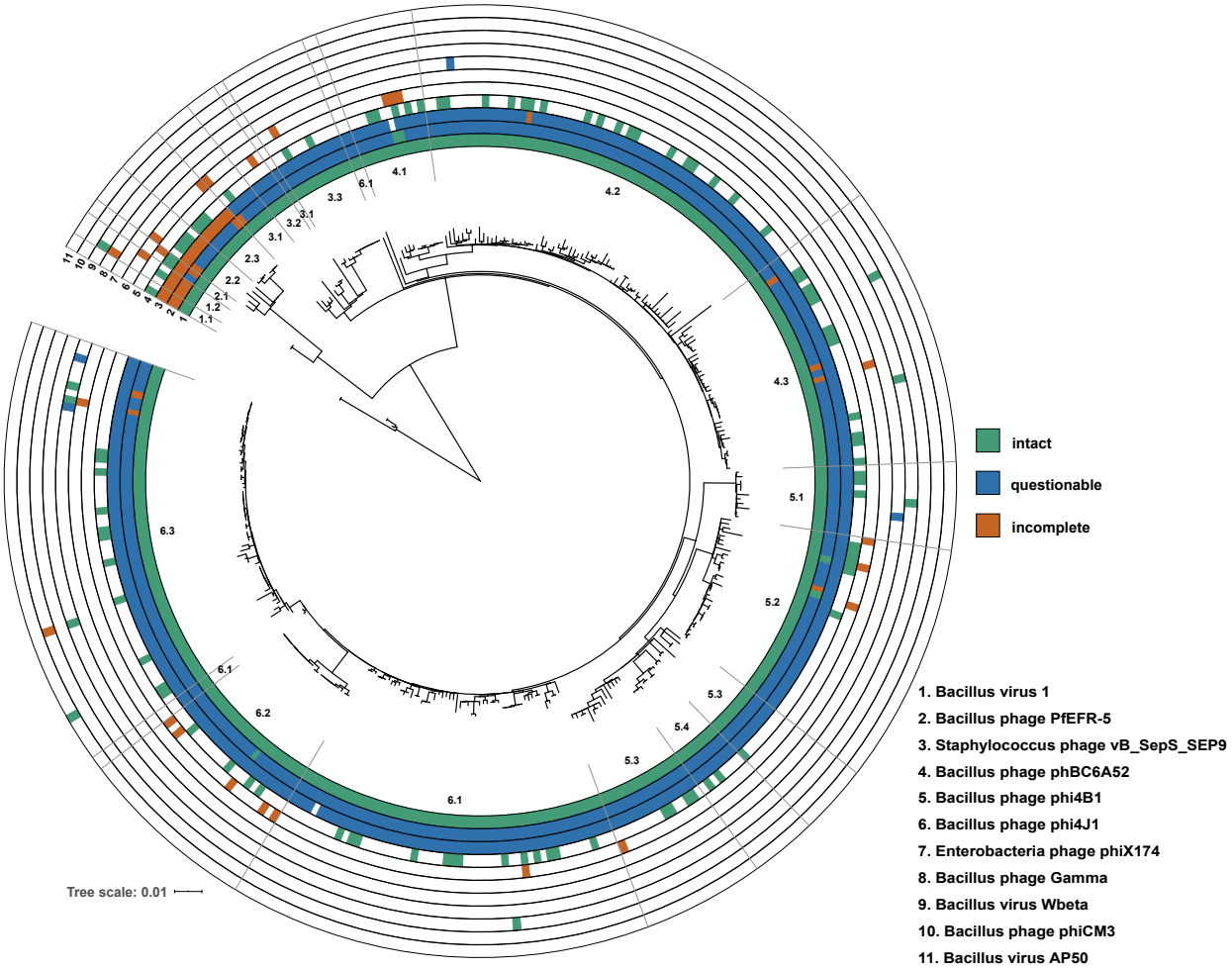

Figure 2. Prophage sequences identified in whole-chromosome tree of 356 global *B. anthracis* isolates. Primary clusters are divided into their numbered nested clades by grey lines. The key on the right indicates the phage sequence present for each isolate, numbered according to their order from the inside of the ring to the outside. Color 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.

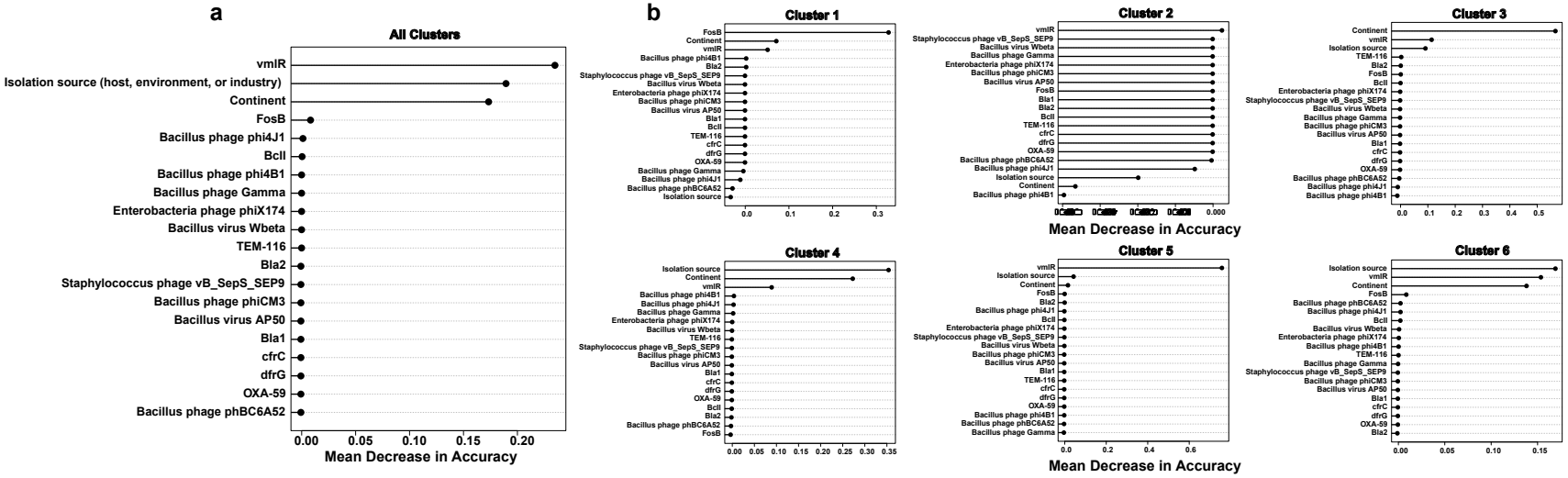

Figure 3. Importance of the covariates in defining population genomic architecture for all primary clusters combined (a) and for each primary cluster on its own (b) by the random forest classifier.