LLNL Genomic Assessment: Viral and Bacterial Sequencing Needs for TMTI, Tier 1 Report

LLNL-TR-417549

September 30, 2009

Correspondent:
Thomas R. Slezak, Associate Program Leader, Informatics
925-422-5746, slezak@llnl.gov

Contributing Authors:
Tom Slezak, 925-422-5746, slezak@llnl.gov
Monica Borucki 925-424-4251, borucki2@llnl.gov
Raymond Lenhoff 925-424-4034, lenhoff2@llnl.gov
Elizabeth Vitalis 925-422-0149, vitalis1@llnl.gov

Chemical and Biological Countermeasures Division
Global Security Program
Lawrence Livermore National Laboratory (LLNL)
Livermore, CA

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

TABLE OF CONTENTS

Executive Summary

We have performed an initial analysis of the sequencing needs for the TMTI program, focusing as directed on the hemorrhagic fever viruses and the Category A bacterial select agents. We have determined that there are multiple knowledge gaps, related to multiple DoD mission needs (including but not limited to TMTI), that need to be addressed by an aggressive genomic sequencing program. These include:

- Strains with observed high or low virulence
- Vaccine or other avirulent strains
- Non-pathogenic (to humans) near-neighbor strains
- Strains with different host range
- Strains with different resistance characteristics
- Strains from poorly-represented geographical regions
- Historic outbreak strains
- Strains that increase coverage of current diversity knowledge

We note that our desires to fill these knowledge gaps generally run into one of three situations when it comes to strain acquisition:

1. The strain is present in a collection somewhere in the US, and presumably is accessible for sequencing via either paid or unpaid collaboration.
2. The strain is known to be in a collection in a foreign country and access is likely problematic at best.
3. The strain is not known to be in a collection, but is presumed to be circulating in the environment in one or more regions.

Even in case (1), there may be significant difficulties involved in obtaining appropriate quantities and quality of nucleic acids for sequencing, even when the collaborator is willing and motivated. This can be due to problems growing sufficient material, regulatory approval delays, contractual delays, etc.

Sequencing Priority Determination

Rather than attempting to present our results as a static list of desired strains, we will instead define a sequencing priority system consisting of a set of three *Sequencing Priority Queues* and general rules for membership on those queues:

- **Priority Queue 0:** *This queue is for operational samples of sufficient priority to DTRA that other work can and should be pre-empted*. This would potentially include outbreak clinical or environmental samples of unknown origin (which are inherently metagenomic sequencing problems and potentially need to be examined for both RNA and DNA pathogen content). It could also include other types of samples determined to be of immediate priority. Currently, we expect that Columbia University will continue to process the priority queue 0 strains that are anticipated to contain RNA viral pathogens. It is less clear at this time where any other types of

operational samples will be processed. We expect this queue to be empty when there is no backlog of current operational samples.

- **Priority Queue 1:** *This queue is for samples that have reasons to be considered to have potential impact on the TMTI countermeasures goal*. This would potentially include strains with metadata indicating unusual virulence (high or low), vaccine strains, avirulent close genetic near neighbors (with "close" being a fuzzy concept), strains with different host range, unusual resistance, etc. The premise for priority queue 1 strains is that their proteomes could be examined by a downstream protein structure pipeline to attempt to determine the structural reasons for the functional differences, thus potentially identifying countermeasure targets. We expect that most isolated viral samples obtained from collaborators will be candidates for this queue, at least until reasonable levels of strain diversity have been sequenced. The TMTI team may evolve a method for prioritizing the members of this queue; this will have to consider the relative importance of bacteria vs. viruses, as well as the relative merits of virulence vs. host range vs. resistance. When there are items on this queue, any TMTI sequencing center needing a work assignment should take items from this queue before taking from priority queue 2.

- **Priority Queue 2:** *This queue is for samples that could help fill phylo-geographical (strain diversity) knowledge gaps, but for which there is currently insufficient associated metadata to warrant placement on priority queue 1*. We would expect the bulk of the various bacterial strain collections now being considered for TMTI sequencing to be on this queue. The TMTI team may continue to evolve a method for prioritizing the members of this queue. One potential option is to screen samples with custom SNP arrays to determine their relative novelty. If a TMTI sequencing center needs a work assignment and priority queue 1 is empty, they will take items from this queue.

This prioritization scheme clearly depends on the available metadata about samples. Much of this will come from the Subject Matter Experts (SMEs) for these organisms, or from the literature. We anticipate that over time, some new information about a sample may cause it to be placed in a different priority queue. As will be detailed in our future Task 1.2 report, we are assuming the eventual development of several TMTI system components to make this priority scheme workable:

- A system for tracking promised strains, receipt of strains, verification and QC of strains.
- A method for determining which strains will be sequenced by which TMTI center
- A system for tracking the sequencing status of all TMTI strains
- A system for the physical receipt/tracking/shipping of strains

Subsequent sections of this report will cover each of the pathogens assigned to us, detailing the relevant knowledge gaps for that pathogen and suggesting potential sources of strains. We have begun initial discussions of access to strains for TMTI sequencing with a variety of key centers (NBACC, CDC, UTMB Galveston, NAU, BYU) and TMTI/Tauri are now beginning the detailed negotiations. As these turn into funded contracts, we anticipate that significant numbers of strains will be obtained. Other strain sources will be examined as needs dictate. We anticipate updating this report to include other important species in the near future.

Introduction:

The Lawrence Livermore National Lab Bioinformatics group has recently taken on a role in DTRA's Transformation Medical Technologies Initiative (TMTI). The high-level goal of TMTI is to accelerate the development of broad-spectrum countermeasures. To achieve those goals, TMTI has a near term need to obtain more sequence information across a large range of pathogens, near neighbors, and across a broad geographical and host range. Our role in this project is to research available sequence data for the organisms of interest and identify critical microbial sequence and knowledge gaps that need to be filled to meet TMTI objectives. This effort includes: 1: assessing current genomic sequence for each agent including phylogenetic and geographical diversity, host range, date of isolation range, virulence, sequence availability of key near neighbors, and other characteristics 2: identifying Subject Matter Experts (SME's) and potential holders of isolate collections, contacting appropriate SME's with known expertise and isolate collections to obtain information on isolate availability and specific recommendations 3: identifying sequence as well as knowledge gaps (eg virulence, host range, and antibiotic resistance determinants) 4: providing specific recommendations as to the most valuable strains to be placed on the DTRA sequencing queue.

We acknowledge that criteria for prioritization of isolates for sequencing falls into two categories aligning with priority queues 1 and 2 as described in the summary. (Priority queue 0 relates to DTRA operational isolates whose availability is not predictable in advance.)

1. Selection of isolates that appear to have likelihood to provide information on virulence and antibiotic resistance. This will include sequence of known virulent strains. Particularly valuable would be virulent strains that have genetically similar yet avirulent, or non human transmissible, counterparts that can be used for comparison to help identify key virulence or host range genes. This approach will provide information that can be used by structural biologists to help develop therapeutics and vaccines. We have pointed out such high priority strains of which we are aware, and note that if any such isolates should be discovered, they will rise to the top priority. We anticipate difficulty locating samples with unusual resistance phenotypes, in particular. **Sequencing strategies for isolates in queue 1 should aim for as complete finishing status as possible, since high-quality initial annotation (gene-calling) will be necessary for the follow-on protein structure analyses contributing to countermeasure development.** Queue 2 for sequencing determination will be more dynamic than queue 1, and samples will be added to it as they become available to the TMTI program.

2. Selection of isolates that will provide broader information about diversity and phylogenetics and aid in specific detection as well as forensics. This approach focuses on sequencing of isolates that will provide better resolution of variants that are (or were) circulating in nature**. The finishing strategy for queue 2 does not require complete closing with annotation.** This queue is more static, as there is considerable phylogenetic data, and in this report we have sought to reveal gaps and make suggestions to fill them given existing sequence data and strain information.

In this report we identify current sequencing gaps in both priority queue categories. Note that this is most applicable to the bacterial pathogens, as most viruses are by default in queue 1.

The Phase I focus of this project is on viral hemorrhagic fever viruses and Category A bacterial agents as defined to us by TMTI. We have carried out individual analyses on each species of interest, and these are included as chapters in this report.

Viruses and bacteria are biologically very distinct from each other and require different methods of analysis and criteria for sequencing prioritization. Therefore, we will describe our methods, analyses and conclusions separately for each category.

Hemorrhagic Fever Viruses:

Collection of data and input from Subject Matter Experts

Tier 1 includes Lassa virus, Marburgvirus, Ebolavirus, Junin virus, Machupo virus,Guanarito virus, , Sabia virus, Chapare virus and Lujo virus. The next tier will include bunyaviruses and flaviviruses and others as TMTI or our team determine to be included.

We carried out extensive literature searches to characterize each virus: ecology, epidemiology including key outbreaks, taxonomy, and genome diversity using phylogenetic trees as available and other resources. This data was analyzed for gaps in sequence knowledge and described within each report. We identified and listed institutions or agencies and relevant scientists as potential sources of isolates that may fill these gaps if sequenced.

Key Subject Matter Experts that were contacted and provided guidance were:

Jim LeDuc, UTMB: Noted that the bunyaviruses are important, and that new emerging viruses have a strong likelihood of arising from this genus. His team has inherited the Bob Shope collection that is an invaluable source of arenaviruses, arboviruses, and bunyaviruses including Rift Valley fever virus.

Tom Ksiazek (UTMB) pointed to the CDC's Stuart Nichol and Pierre Rollin as sources of partially sequenced Lassa viruses and also noted that reservoirs for the viral hemorrhagic fever viruses are not always known. Research in this area is needed.

CDC has numerous branches that require genomic sequencing. Fortuitously, we have had a sustained collaboration since 2003 with Mike Frace who coordinates all sequencing efforts at the CDC. We have arranged for negotiations between TMTI and CDC, and once those discussions have begun, we hope to be able to obtain input directly from our numerous CDC contacts. Mike has asked us to hold off on that until the high-level discussions can take place and reach at least a preliminary agreement.

A migratory bird group from UCLA Center for Tropical Research (CTR) has an immense sample set of 90K avian cloacal and blood samples that may be extremely valuable for genetic shift and drift studies. In addition, they have numerous samples containing arboviruses, avian malaria, and other viruses covering a broad geological and host range. We have had discussions with CTR Director Tom Smith and members John Pollinger and Emily Curd, and they would be willing to work with TMTI as deemed appropriate. We note that studies of migratory birds present one of the few opportunities to sample certain classes of pathogens circulating in countries that are recalcitrant to release samples to the US.

Samples that may illustrate a viral species jump have been identified at the California Department of Public Health in Richmond, CA. Sharon Messenger has a large set of rabies virus samples that appear to

represent a rare species jumping event from striped skunks to gray foxes. Further details are in Appendix I. In addition, a sequencing whitepaper will be submitted by Sharon and members of our team to propose a study of the viral evolution mechanisms underlying this apparent species jump.

Conclusions: Current status and specific suggestions for sequencing of hemorrhagic fever viruses

By default, most viral hemorrhagic fever samples will be considered Tier 1. We have noted that there is a predominance of sequence data from human isolates, yet animals are the reservoirs for most viruses. Therefore, we recommend increasing host range knowledge for most of the viruses via sequencing focused on reservoir species.

The individual reports describe our findings in the following areas for each hemorrhagic fever virus:

- Human epidemic coverage
- Vector/reservoir species coverage
- Geographical and temporal coverage
- Phenotype (virulence coverage)
- Near neighbors coverage
- Potential sources of samples

Lassa virus

- Lassa virus causes thousands of deaths each year in West Africa and is the most common imported viral hemorrhagic fever agent in Europe and the United States.

- Lassa virus strains are antigenically and genetically diverse and may differ in nucleotide sequence by up to 27%. Genetic distance correlates with geographic distance.

- Seroprevalence studies and risk map analysis indicate that Lassa virus is also endemic in countries with no reported cases of Lassa fever, including Ivory Coast, Ghana, and Benin. The identification of seroprevalence in these countries is not surprising as the range of the rodent vector of Lassa virus extends beyond the endemic region.

- Ten Lassa virus strains have been sequenced, however, critical data such as date of isolation, host, and location are not available for many of these strains, and all may be human isolates. Suggested strains for whole genome sequence analysis are:

    - Rodent vectors: Recent and historical isolates from the principal host, *Mastomys natalesis* should be sequenced. An effort should be made to sequence isolates collected from throughout the endemic region. If possible, isolates from rodents outside the endemic region should be collected and sequenced as well.

    - Human hosts: Geographically and temporally diverse isolates from obtained from mild and severe (hemorrhagic and/or neurologic) from various regions within the endemic area should be sequenced. If possible, human isolates from outside the endemic region should be collected and sequenced as well.

- Genetic near neighbors:

    - Mopeia virus: The two strains that have been sequenced appear to be diverse, more strains should be sequenced to allow better a comprehensive view of species diversity.
    - Mobala virus: One strain has been sequenced; more strains should be sequenced to allow better a comprehensive view of species diversity.
    - Ippy virus: Serological studies indicate that strains are antigenically diverse, however only one strain has been fully sequenced. Additional isolates should be sequenced to allow a more comprehensive view of species diversity.

Marburgvirus

- Whole genome sequence data is available from human isolates from all marburgvirus epidemics, however, strains of marburgviruses isolated from the same outbreak may exhibit high overall genetic diversity, and therefore multiple isolates should be sequenced from each outbreak.

    - Eleven human isolates were sequenced from the 2004-2005 Angola outbreak, however only the three isolates sequenced from the 1998-2000 outbreak in the Democratic Republic of Congo. The three strains were found to be genetically diverse; therefore more isolates from this outbreak should be sequenced. Ideally, the isolates should be selected to vary geographically, temporally, and in clinical outcome.

- Whole genome sequence data is available from one isolate from one of the three species of bats that serve as a marburgvirus vector.

    - Potentially infected bat tissue samples should be screened for the presence of viral RNA and an effort should be made to obtain the entire genome sequences from multiple samples from all three of the known vector bat species.

- Ecological modeling studies and evidence of marburgvirus infection in bats indicate that marburgvirus is present in countries that have not yet reported cases of Marburg hemorrhagic fever such as Tanzania, Ethiopia, and Zambia; potentially infected samples from these regions should be included in the study if possible.

- Near neighbors are members of the *Ebolavirus* genus; see next section for ebolavirus sequencing priorities.

Ebolavirus

- Whole genome sequence data is needed from human and animal hosts and vectors.

    - There is no whole genome sequence data available for 8 of the 18 ebolavirus outbreaks. Ideally several isolates from each large outbreak should be sequenced to ascertain genetic variability, especially in the cases where multiple introductions are likely (for example, when the affected population was exposed to a large colony of bats). The isolates should be selected to vary geographically, temporally, and in clinical outcome.

    - There is no whole genome sequence data from infected bat and ape samples. Potentially infected bat and ape tissue samples should be screened for the presence of viral RNA and an effort should be made to obtain the entire genome sequences from multiple samples from animal hosts in a variety of geographical locations and years.
    - Although limited data from the 1996 Gabon outbreak indicate that the genotypes of viruses obtained from survivors, fatal cases, and asymptomatic cases do not appear to differ; this should be confirmed by analysis of relevant isolates from other ebolavirus outbreaks.
- Ecological studies suggest that ebolavirus is present in countries that have not yet reported cases of ebolavirus hemorrhagic fever (such as Cameroon and Ghana); potentially infected samples from these regions should be included in the study if possible.
- Near neighbors are members of the *Marburgvirus* genus; see previous section for *Marburgvirus* sequencing priorities.

Junin virus

- Although a large number of Junin virus strains have been isolated from humans and rodents, only four strains have been fully sequenced, all of which are laboratory-passaged strains obtained from humans decades ago. Additionally, three of the four strains sequenced are from derived from the same isolate.

    - Whole genome analysis is needed for viral strains isolated more recently from humans and from rodent vectors from different geographical regions of the endemic area.
    - Isolates from both mild and severe cases of Argentine hemorrhagic fever should be sequenced.
    - Additional isolates of near neighbors, Machupo virus (see next section for sequencing priorities) and Tacaribe virus should be sequenced.

Machupo virus

- Whole genome sequence data exists for two strains of Machupo virus, and both of these strains are from human cases that occurred decades ago. Therefore, suggested strains for whole genome sequence analysis include isolates obtained from more recent human cases as well as isolates obtained from the rodent vector (vesper mouse).
- Additional isolates of near neighbors Junin virus (see previous section for sequencing priorities) and Tacaribe virus should be sequenced.

Guanarito virus

- The only fully sequenced genome of Guanarito virus is an isolate from a 1990 human case of Venezuelan hemorrhagic fever, therefore temporally, geographically, and genetically diverse human isolates need to be sequenced.
- The rodents that vector Guanarito virus have a distribution much broader than that of Venezuelan hemorrhagic fever, and infected rodents have been trapped outside of the Venezuelan hemorrhagic fever endemic region. Temporally and genetically diverse isolates obtained from infected rodents found inside and outside of the endemic region should be sequenced.

- If possible, isolates of near neighbors Amapari, Cupixi, Tacaribe, and Sabia viruses should be sequenced as one completely sequenced genome exists for each of these viruses. However, it is likely that no additional virus isolates are available for one or more of these near neighbors.

Hemorrhagic fever-associated arenaviruses for which the animal vector is unknown: Lujo virus, Chapare virus, and Sabia virus

- Three arenavirus species, Lujo virus, Chapare virus, and Sabia virus, have been isolated from human hemorrhagic fever cases but have not yet been detected in an animal vector, therefore the geographical range and ecology of these viruses remains undetermined and the potential for future outbreaks cannot be predicted.
    - Lujo virus was identified in South Africa in 2008 as the cause of a small outbreak of hemorrhagic fever that originated in Zambia. This outbreak involved secondary and tertiary spread and had a mortality rate of 80% (4/5 cases).
    - Chapare virus was isolated from a fatal case of hemorrhagic fever that was part of a small outbreak that occurred between December 2003 and January 2004 near Cochabamba, Bolivia, however political issues may make the region difficult to access.
    - Sabia virus emerged in 1990 when it caused a fatal case of hemorrhagic fever that occurred in Sao Paulo, Brazil. Subsequently the virus caused nonfatal influenza-like illness in two laboratory technicians who were exposed while working with the virus.

- Genetic and ecological knowledge for each of these viruses is limited to that obtained from a single human outbreak. The small number of clinical samples available for genetic analysis limits the robustness of assays developed for detection of these viruses.

- Whole genome sequence data is needed for the near neighbors of Sabia virus, Chapare virus, and Lujo virus including Cupixi virus and Ippy virus. However, it is likely that only a single isolate exists for each of these viruses as there was only one strain identified in searches of scientific literature and in Genbank.

Category A Biothreat Bacteria:

Collection of data and input from Subject Matter Experts

Burkholderia pseudomallei and *Burkholderia mallei, Francisella tularensis, Bacillus anthracis, Yersinia pestis, Brucella species.*

Next tier may include *Clostridium botulinum, Salmonella typhi, Rickettsii* species, *Vibrio cholera*, and others as TMTI or we determine to be included.

Thorough literature searches were used to characterize each of the five Category A bacterial biothreat agents specified by TMTI for our initial focus. For each agent, information regarding the ecology (including geolocation of sequenced isolates when known), epidemiology, taxonomy, genome diversity using phylogenetic trees (generated in house and others available from the literature), genetic near

neighbors, as well as newly emerging strains and antibiotic resistance was collected. This data was analyzed for gaps in sequence knowledge and described within each report. We identified and list institutions or agencies and relevant scientists as potential sources of isolates that may fill these gaps in sequence.

Input was sought and obtained on all five bacterial agents from experts on each species from Dr. Paul Keim's lab at NAU:

They have extensive isolate collections and are pursuing funds to have these sequenced. They have written sequencing proposal whitepapers for each of the major bacteria. These provide additional background information regarding biology, genetics, epidemiology, phylogeny, and geographical distribution (from their knowledge base) as well as specific suggestions for sequencing based on their collections. In addition, they provided answers in e-mail dialogue to several questions we had regarding the bacterial pathogens from their perspective. The input from the Keim lab is extremely valuable yet ought to be viewed in the context of DTRA objectives (we note that not all their recommendations may be applicable to TMTI goals). The focus on forensics research in that lab has given valuable perspective of sequencing needs for accurate detection and forensics, but is generally less focused on aiding countermeasure development.

Appendix II contains the Keim lab whitepapers as well as a document with a compilation of their e-mail responses to our questions asked in relation to their draft reports. The whitepapers have been edited in response to some of our questions since we sent drafts to Greg Myers on 9/18/09. A high-level summary includes:

1. Ancient geographical spread of *Bacillus anthracis:* sequencing suggestions to understand geographical source(s) of ancient strains.
2. Population genetics and evolution of *Burkholderia pseudomallei:* sequencing suggestions to better understand its evolution.
3. Filling in the gaps in the worldwide distribution of *Brucella:* sequencing suggestions to understand distribution and diversity of *Brucella.*
4. *Francisella tularensis:* sequencing suggestions to better understand phylogeny.
5. Whole genome sequencing needs in *Yersina pestis:* sequencing suggestions to better understand diversity. A pointed suggestion is to sequence numerous strains in the ORI group, as it is widespread geographically and highly virulent.

Rich Robison at BYU was consulted and is generously open to sharing information. He houses a large isolate collection of ~1500 isolates, largely obtained under DHS funding:

~300 *Burkholderia* species strains (mostly *B pseudomallei* except ~20 *B. mallei*); most of these are obtained from Tyrone Pitt and Bart Curry and are from Australia. In addition, have 10 *Burkholderia andropogonis,* an important near neighbor.
~300 *Bacillus anthracis*
~300 *Yersina pestis* mostly obtained from Health Departments in New Mexico and Utah
~100 *Francisella tularensis* mostly obtained from Health Departments in New Mexico and Utah
~120 *Brucella* species

For each isolate the Robison lab carries out a systematic three-tiered analysis to characterize the isolate and writes a standardized Certificate of Analysis (COA). He is willing to provide these for our perusal to help identify potential candidates for sequencing.

Jim Burans of NBFAC has numerous bacterial isolates as part of the DHS strain collection.

This collection is currently housed elsewhere (BYU, Battelle, etc.) while the new NBFAC building is in the process of being certified for lab operations. NBFAC is already providing some isolates to ECBC to sequence and we anticipate this will continue. NBFAC has suggested to DHS that TMTI could potentially assist in sequencing some of the *Burkholderia pseudomallei* isolates that will be part of an international collaboration with Australia. We anticipate increasingly closer relationships between TMTI and NBFAC.

UTMB has bacteria expertise and isolates, particularly *Yersinia pestis.*

We recently received the list of bacterial strains that NMRC has completed and in progress, and are pleased to note that many of the gaps we identified are being filled, especially genetic near neighbors of *Yersinia pestis*, *Bacillus anthracis,* and targets and near neighbors for *Burkholderia pseudomallei.*

We anticipate that other gaps will also be filled by the recent (unpublished) and ongoing sequencing efforts at LANL and ECBC.

Members of the USDA National Veterinary Services lab were contacted regarding animal isolates and vaccine strains. Recommendations for *Bacillus anthracis, Burkholderia, Francisella tularensis,* and *Brucella* were gathered from Dr. Matt Erdman and Kristina Lantz and the e-mail dialogue is in Appendix III.

Conclusions: Current Status and Specific Suggestions for Bacterial Sequencing. We have listed the 5 Category A bacteria in the order of relevance to TMTI objectives according to our analyses.

Known virulence and antibiotic resistance potential are primary criteria for inclusion in queue 1, and therefore, we have extracted what is known in these categories for the five bacteria. We note that for the five bacteria in this report, antibiotic resistance strains are not as abundant as they are for other bacteria such as *Staphylococcus aureus* and *Mycobacterium* species. Therefore, when identified or discovered, antibiotic resistant strains of these bacteria ought to be top priority for sequencing to aid in the characterization of antibiotic resistance mechanisms. In addition, for top priority consideration are attenuated strains of virulent counterparts, as the comparison of their sequence will aid in the determination of molecular determinants of virulence and thus potential countermeasure targets.

Sequence of genetic near neighbors is important to understand virulence of pathogenic bacteria, as well as to aid in understanding phylogeny. We have included suggestions for sequencing near neighbors as we have identified gaps; however, note that additional significant near neighbors may be discovered for any of the bacterial species that may become a high sequencing priority, especially if they have particularly interesting virulence (or lack thereof), resistance, or host traits. The TMTI priority process must achieve a balance between static strain collection lists and dynamic discovery.

Burkholderia pseudomallei* and *Burkholderia mallei

A large amount of unknown genetic variation for *B. psuedomallei* due to newly discovered pathogenicity island differences combined with multiple mechanisms of antibiotic resistance make sequencing additional genomes of high importance.

B. mallei is a human pathogen of poorly understood virulence due to the scarcity of natural human infections. In addition, only 11 complete genome sequences of *B. mallei* exist, four of which are derived from a single infection.

Queue 1

- There is a need for genomes of isolates of *B. pseudomallei* from non-human hosts.
- Additional isolates from Thailand (source of isolates associated with human disease) are needed to better define the spectrum of *B. pseudomallei* human virulence.
- There are currently no avirulent or vaccine strains of either *B. pseudomallei* or *B. mallei,* and therefore, if any should be discovered, they ought to be top priority to help understand virulence mechanisms in these bacteria.

Queue 2

- Additional *B. mallei* genomes should be sequenced, especially from S. America and Africa.
- Genomic sequence from near neighbors other than *B. cepacia* is needed.

DHS has initiated an international collaboration with Australia. They will require assistance sequencing up to 90 isolates of *B. pseudomallei* over the next 2 years. The plan is to prioritize ~100 isolates for sequencing using a *B. pseudomallei* SNP microarray to be customized by LLNL. NBFAC plans to sequence 10 of the isolates. Both clinical and environmental isolates will be sequenced. Details are still being worked out; however DHS may wish to enlist help from TMTI to sequence many of these strains

Francisella tularensis

Unknown environmental reservoirs of this bacteria and poorly defined mechanisms of virulence make this a priority for additional focused genome sequencing.

Queue 1

- *F. tularensis* subsp. *mediasiatica* (*F.t. mediasiatica*) virulence is uncertain due to the scarcity of isolates to date. Whole genomic sequencing is needed for the subspecies *F.t. mediasiatica*, particularly from Japan.
- Any avirulent and/or vaccine strains of *F. tularensis* that may be discovered (in addition to the LVS strain) should be top priority to help understand virulence mechanisms in these bacteria. Amy Rasley at LLNL has data supporting that genomic sequence for 7 of her clinical isolates may be valuable to better understanding *F. tularensis* virulence. See Appendix IV for details.
- Any strains exhibiting antibiotic resistance ought to be sequenced, as very little is known about potential antibiotic resistance mechanisms in these bacteria.

Queue 2

- Additional *F.t. holartica* isolates from Japan should be sequenced, and *F.t. holartica* isolates from California subclade B.br.002/003 should be sequenced.

- *F.t. novicida* isolates from Australia that are known to cause human infection should be sequenced.
- Near neighbors *Wolbachia* species and *F. pisicicida* should be sequenced.

Bacillus anthracis

The spore stability of this species keeps it near the top of the bacterial agents despite an abundance of existing sequence data and reduced antibiotic resistance mechanisms when compared to gram negative bacteria. Due to lack of solid correlation between a particular genotype and human pathogenicity, studies of pathogenicity have focused on the plasmids that confer virulence such as the capsule and edema factor, lethal factor and protective antigen.

Queue 1

- To better understand virulence, additional isolates *of B. cereus* obtained from human inhalational anthrax cases should be sequenced.
- Unusually virulent strains or alternatively, avirulent and/or vaccine strains ought to be sequenced if they should become available (ex vaccine strains: *B. anthracis* Carbosap and *B. anthracis* strain 55 (see NVSL input)).

Queue 2

- Additional *B. anthracis* isolates from Africa should be sequenced.
- Additional *B. anthracis* isolates form N. America that comprise the C group should be sequenced.
- Near neighbor sequence is needed, especially, *B. mycoides and B. weihenstephanensis.*

Yersinia pestis

Given the extensive existing data for this species, and the information from NAU that 160 new genomes from China are being sequenced, and will be publicly released in the near future, it has a lower priority for TMTI's sequencing program than the above bacteria.

Queue 1

- To better understand virulence, isolates from *Y. pestis* Pestoides and *Y. pestis* biovar Microtus genomes (generally avirulent to humans but virulent in rodents) should be sequenced.
- Very little is known about antibiotic resistance in this bacteria, and therefore, should any resistant strains be discovered, these ought to be top priority for sequencing.
- Highly virulent strains and avirulent strains (if they can be obtained) from the ORI molecular group may provide valuable information about virulence in *Y. pestis.*

Queue 2

- Isolates of *Y. pestis* that are similar to *Y. pestis* Angola should be sequenced due to their diversity.
- To broaden the scope of near neighbors, *Y. frederickensii, Y. rohdei,* and *Y. ruckeri* and others should be sequenced.

***Brucella* species**

To date, *Brucella* has lacked acute extreme virulence, and thus is our lowest priority for Category A bacterial sequencing for TMTI's needs.

Queue 1

- Little is known about its virulence or antibiotic resistance mechanisms, particularly due to the lack of strains sequenced from human infections. There are strains showing antibiotic resistance (for ex*ample B. abortus* strain RB51; see NVSL input), and these should be top priority for this organism in addition to pathogenic strains isolated from humans.

Queue 2

- Additional human isolates of *Brucella melitensis* (*B. melitensis*), *Brucella suis* (*B. suis*) and *Brucella abortus* (*B. abortus*) should be sequenced.
- Additional human isolates from newly discovered *Brucella* species, *Brucella pinipedialis, Brucella ceti,* and *Brucella inopinata* should be sequenced.
- For complete phylogenetic data, isolates from biovar 5 of *B. abortus* as well as biovars 1, 2 and 4 of *B. suis* should have at least one complete genome sequenced.
- *Brucella* from the Middle East (especially *B. melitensis*) should be sequenced if available.

Next steps for selecting isolates for sequencing

TMTI-wide Strain Tracking

We have listed in the Executive Summary the key organizations that could potentially provide access to many of the strains needed to fill the various gaps listed above. Each organization has their own unique motivation for potential collaboration and their own sets of desired conditions for useful collaboration to actually occur. It is almost certain that actual access to strains will occur in unpredictable order; our suggestion for priority queues 1 and 2 reflect this condition. Most strains received will likely not have known phenotypical traits that could potentially influence countermeasure design; hence they will go on priority queue 2. Access to strains from recalcitrant parts of the globe will likely prove to be both rare and difficult/expensive to obtain. *In some cases, it may prove advantageous to sponsor indigenous sequencing of important samples that cannot otherwise be exported to the US.*

Regardless of the strain acquisition difficulties, there is a need for TMTI to establish centralized systems and procedures to:

- Track samples potentially available, desired, or promised by collaborators
- Track samples received
- Track metadata about samples
- Track data used to screen samples (e.g., microarray data)
- Track sample priority decisions
- Track sample disposition and subsequent sequencing status

Note that this need is a separate system than the LIMS systems that are needed to track the activities within each individual TMTI-sponsored sequencing center. This will be discussed in more detail in our forthcoming Task 1.2 report.

Strain Screening: Which strains warrant sequencing, and in what order?

Access to strains tends to follow a feast-or-famine pattern. In general, the relatively few strains that are clearly destined for priority queue 1 are difficult to obtain. However, access to relatively large strain collections presents a different problem: determining which strains are most worthy of the time and expense of sequencing, in some priority order. LLNL has been developing and testing a set of microarrays for NBACC that focus on both SNPs and the presence/absence of genes for the Category A bacterial agents. Originally designed for rapid forensic determination, their low cost and accurate performance has convinced NBACC that they also could be used for several other missions, including:

- Verify strain collections (ensure that any major changes due to passage are highlighted).
- Allow rapid typing of unknown isolates, to determine their relative positioning compared to all other known/typed isolates. For example, a new *Y. pestis* strain that is only 10 SNPs and 2 genes different from the Kim strain will be a lower priority to sequence than a strain that is more than 200 SNPs and 40 genes different from any strain currently typed. (Note that if the new strain close to the Kim strain had an unusual phenotype, it should already be destined for high-priority sequencing on priority queue 1).
- Determining that two or more strains in a newly-received collection are essentially equivalent and thus only one of the set should be sequenced (if any). We have seen that many collections of unknown strains contain exact or near-duplicates, which otherwise is not discovered until a $10K 454 run has been completed and assembly and analysis performed.

LLNL has also developed another microarray, the Microbial Detection Array (MDA) that complements the SNP/gene array. The MDA array is being used to support NBACC in a variety of ways that may be pertinent to the TMTI sequencing effort, including:

- Provide bacterial and viral metagenomic readout of a sample. This can be used to identify sample contamination or mis-labeling issues.
- For organisms without a SNP/gene chip available, it provides a way to map the “distance” of a sample to all other strains of that organism for which either whole genome sequence or a prior MDA sample are available. This can be used to prioritize for sequencing one or more new samples compared to the available known strains. As noted for the SNP chip, it can also highlight extremely close unknown strains, to avoid duplicate sequencing.

To maximize the value of TMTI sequencing dollars, we suggest that these sorts of sample screening procedures be considered to prioritize samples. Other sources of relevant arrays may be available.

Detailed organism reports:

Lassa virus: Lassa Fever

Summary

Lassa virus is the causative agent of Lassa fever, a severe hemorrhagic fever that causes thousands of deaths each year in West Africa. Lassa virus is the most common imported viral hemorrhagic fever agent in Europe and the United States. Lassa virus strains are antigenically and genetically diverse and may differ in nucleotide sequence by up to 27%. Currently ten Lassa virus strains have been sequenced, however, critical data such as date of isolation, host, and location are not available for many of these strains, and all may be human isolates. Therefore, it is important to sequence a comprehensive and diverse set of human and rodent isolates in order to design assays capable of detecting genetically heterologous Lassa virus strains associated with human disease. This has been recently illustrated by the recent discovery of a new Lassa virus strain that appears to represent a new lineage within the species.

Eight new arenavirus species have been discovered in the last five years, three of which cause hemorrhagic fever in humans. One of the newly detected hemorrhagic fever arenavirus species, Lujo virus, was discovered in South Africa in 2008 and had a mortality rate of 80% [1]. It is likely that additional genetically distinct pathogenic arenaviruses remain to be discovered in Africa and elsewhere.

Ecology and epidemiology

Lassa fever results in an estimated 5,000 deaths annually [2]. Although most Lassa virus infections are asymptomatic, the health burden associated with Lassa virus is significant; about 30% of all adult deaths in endemic regions are caused by Lassa virus. The mortality rate for Lassa fever is 15-20% in hospitalized cases and case fatality can reach 50% during outbreaks. Mortality rate is very high for pregnant women and their fetuses, especially during the third trimester when the disease results in death of the mother and/or fetus in greater than 80% of the cases [3].

Recent outbreaks of Lassa fever include an outbreak that occurred in Sierra Leone from January 1996 to April 1997 and resulted in 799 hospitalized cases and 148 deaths; and a 2004 outbreak which occurred in the pediatric ward of a Sierra Leone hospital and resulted in 95 pediatric cases with a case fatality rate of 30-50% in patients under five and 71% in those less than one year of age [4]. This outbreak also claimed the life of Lassa fever specialist Dr. Aniru Conteh.

Lassa virus is the most common imported hemorrhagic fever virus in Europe and the United States. In year 2000, four Lassa virus fever cases were imported into European countries, and each of these infections proved to be fatal [5]. In fact, the newly discovered AV strain was isolated from a German tourist who had visited Burkina Faso, Ghana, and the Ivory Coast, three West African countries that had never reported a case of Lassa virus hemorrhagic fever [3].

Lassa virus is present in West Africa and clinical infections are regularly reported from Nigeria, Sierra Leone, Liberia, and Guinea (Figure 1A). Seroprevalence studies indicate that Lassa virus is also endemic countries with no reported cases of Lassa fever, including Ivory Coast, Ghana, and Benin, with a seroprevalence of 20% detected in Ivory Coast [6]. The identification of seroprevalence in these countries is not surprising as the range of the rodent vector of Lassa virus extends beyond the endemic region (Figure 1B).

Multimammate rats (*Mastomys natalensis*) are the natural host of Lassa virus and may also transmit three closely related nonpathogenic arenaviruses, Mopeia virus, Mobala virus, and Ippy virus [7]. These rodents are ubiquitous in sub-Saharan Africa, frequently enter human dwellings, and shed the virus in their excreta. Humans become infected by contact with infected rat tissue (multimammate rats serve as

a food source in Lassa endemic areas), contamination of food with rat excrement, or by inhalation of aerosolized virus from rat excrement. Human-to-human transmission may also occur.

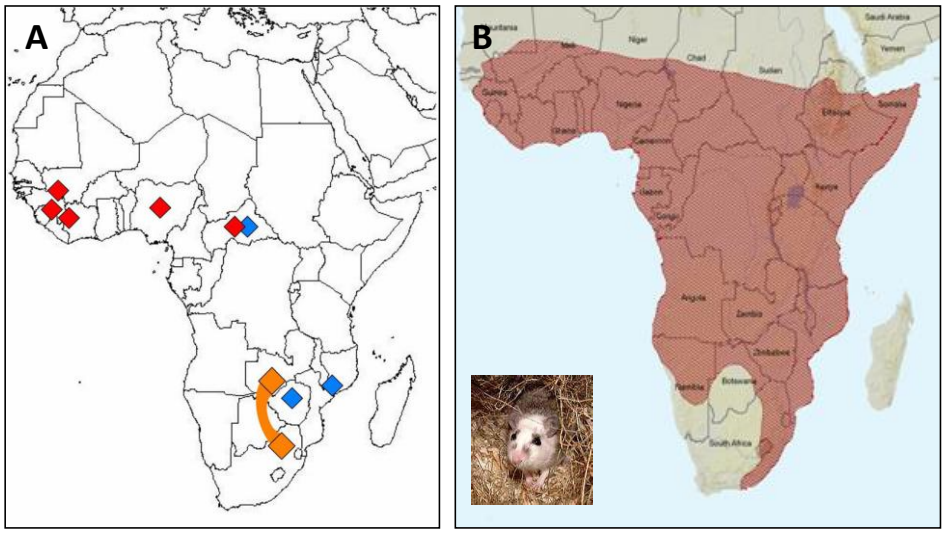


Figure 1. Comparison of the geographic range of African arenaviruses with that of the Lassa virus rodent vector. (A) Geographic distribution of African arenaviruses. Mobala virus, Mopeia virus, and Ippy virus (blue) have not been implicated in human disease; Lassa virus (red) can cause hemorrhagic fever. The origin of the newly discovered arenavirus, Lujo virus, index and secondary and tertiary cases linked in the 2008 outbreak are indicated in gold [1]. (B) The geographic distribution of *Mastomys natalensis,* the rodent vector of Lassa virus.

As is the case with seroprevalence data, risk map analysis predicts that Lassa virus should be present in other, nonendemic regions of West Africa, with the regions between Guinea and Cameroon representing the greatest risk (Figure 2) [7]. Additionally, risk map analysis identified favorable environmental conditions, primarily annual rainfall amount, in several other African countries such as Central African Republic and Tanzania. The authors of the study hypothesized that infection of the rodent vector of Lassa virus (*Mastomys natalensis*) with other Lassa-like viruses (Mopeia virus, Mobala virus, and Ippy virus) may restrict Lassa virus from becoming endemic in these areas. Indeed, a new Lassa-like virus, Morogoro virus, was recently isolated from *Mastomys* species in Tanzania [8]. Alternatively, genetic differences between the subpopulations of the rodent vector species may restrict Lassa virus infection to certain reproductively-isolated subpopulations ("cryptic species") [9].

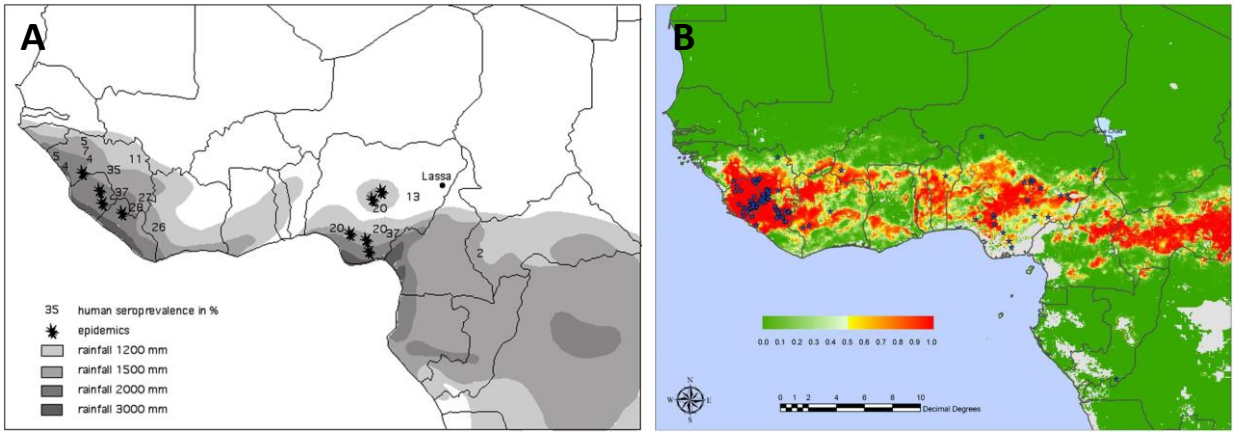


Figure 2. Comparison of Lassa virus outbreaks and seroprevalence with the predicted geographic distribution of Lassa virus based on risk analysis. (A) West and Central Africa mean annual rainfall (1951–1989), Lassa fever nosocomial outbreaks (stars)

and human seroprevalence (numbers in %). (B) The mean predicted Lassa risk map for West Africa with positive localities indicated by stars. Risk map: the posterior probability color scale, from 0.0 (no risk) to 1.0 (highest risk) is shown as an inset. Grey areas are either areas with no suitable imagery (because of cloud contamination; coastal Nigeria and Cameroon) or else are so far from any of the training set sites in their environmental conditions that no predictions are made for them [7].

Viral taxonomy and genome diversity

Lassa virus is a member of the *Arenaviridae* family, which includes at least 22 recognized species and two serocomplexes; eight new arena viruses have been discovered in the last five years [1, 8, 10, 11]. Both serocomplexes include hemorrhagic fever viruses; the Old World serocomplex includes Lassa virus and newly discovered Lujo virus, and the New World serocomplex includes Machupo virus, Junin virus, Sabia virus, Guanarito virus, and newly discovered Chapare virus (Figure 3A). The arenavirus genome consists of two single-stranded segments of RNA, the S segment (3.4 kb) and the L segment (7.2 kb). Both segments are arranged in an ambisense orientation (coded for in opposite polarities).

Lassa virus isolates are a genetically, antigenically, and geographically diverse (Figure 3B)[12]. Genetic distance correlates with geographic distance (Figure 4A), but not with temporal distance, although lack of correlation with temporal distance may be due to the limitations associated with obtaining isolates from historical outbreaks [2]. Strains may differ in nucleotide sequence by up to 27% and up to 15% in amino acid sequence. Five phylogenetic lineages have been proposed [13].

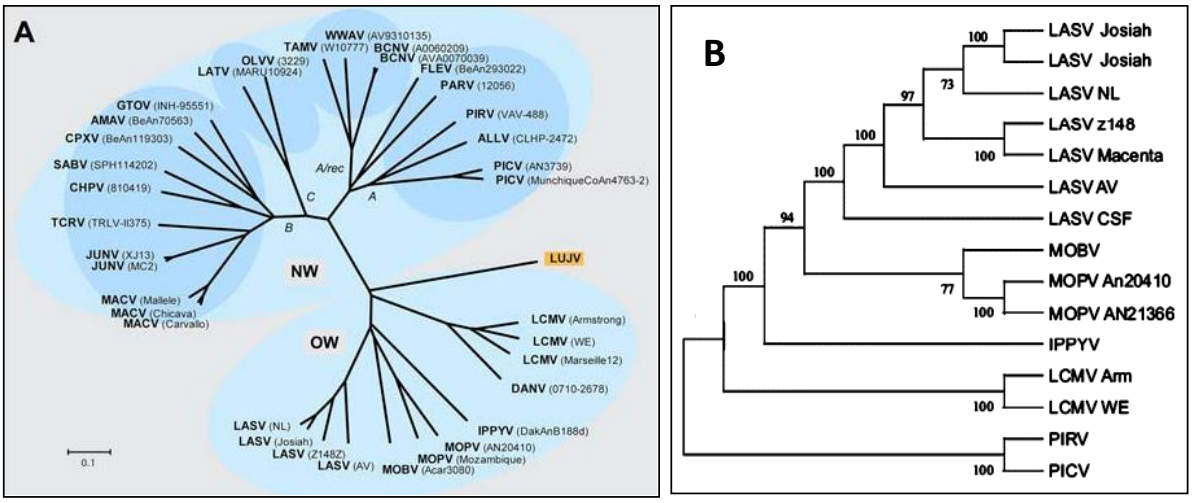


Figure 3. Phylogenetic comparison of Lassa virus strains with other members of *Arenaviridae*. (A) Phylogeny of Old World (OW) and New World (NW) arenaviruses based on the analysis of complete sequences of S segment [1]. South American hemorrhagic fever viruses Junin (JUNV) and Machupo (MACV) show less genetic diversity as compared to Lassa virus(LASV) strains. Newly discovered African hemorrhagic fever arenavirus, Lujo virus, branches off the root of the OW arenaviruses, and may be a novel genetic lineage, separate from the lymphocytic choriomeningitis virus (LCMV) lineage which includes LASV, Mopeia virus (MOPV), and Mobala virus (MOBV). (B) Phylogeny of Old World arenaviruses based on sequence analysis of the NP gene. Sequences from NW arenaviruses Pirital (PIRV) and Pichinde (PICV) were used to root the tree [8]. Whole sequence data is available for each of the Lassa virus strain shown in this figure.

Lassa virus differs in amino acid sequence from near neighbor Mopeia virus by 23-26% [14]. Mobala virus is a nonpathogenic arenavirus and differs from Lassa virus only slightly more as compared to Mobala virus (about 1% more divergent) [2]. Morogoro virus differs from Mopeia virus by 11-26% (amino acid sequence) (Figure 3B) [8]. Ippy virus differs from Lassa virus by 31.6% amino acid sequence (Supporting table S1) [1].

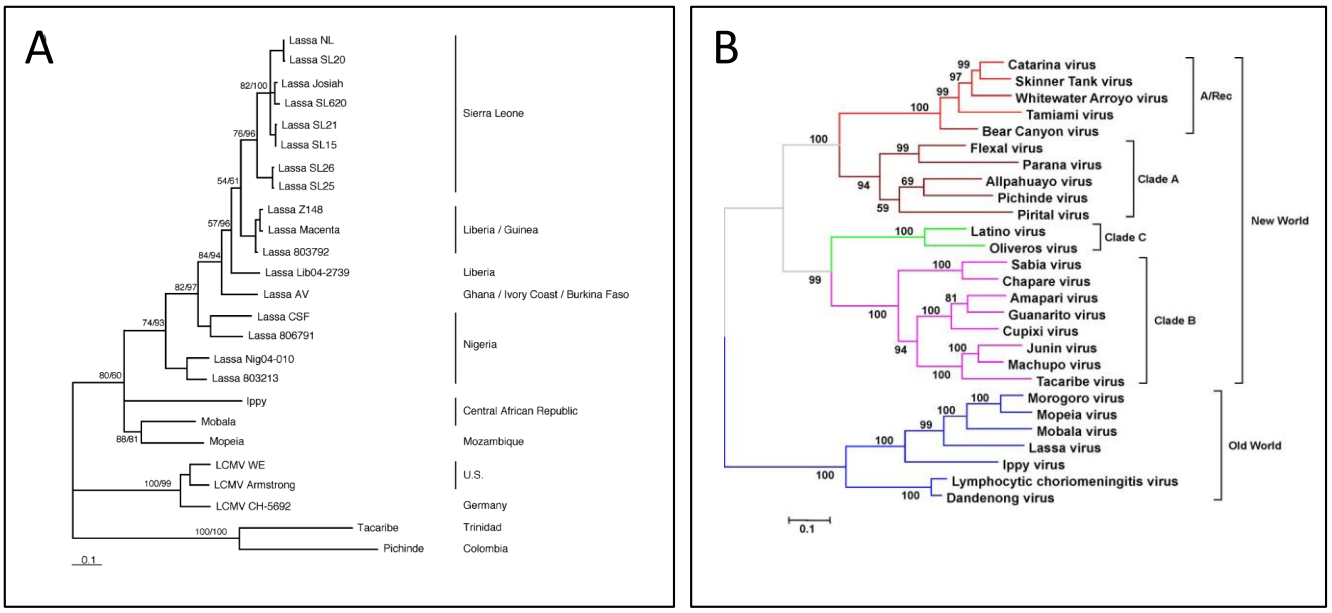


Figure 4. Phylogenetic relationships among arenavirus species based on partial L gene sequences [15, 16]. This dendrogram illustrates correlation of phylogeny with geography as well as the genetic diversity within the Lassa species.

Gaps in sequence data

A relatively large number of strains have been isolated from humans and rodents [2, 9], however from the information available, it appears that only human isolates have been fully sequenced.

Suggested strains for whole genome sequence analysis are:

1. **Rodent vectors**: Recent and historical isolates from the principal host, *Mastomys natalesis* should be sequenced. An effort should be made to sequence isolates collected from throughout the endemic region. If possible, isolates from rodents outside the endemic region should be collected and sequenced as well.
2. **Human hosts**: Geographically and temporally diverse isolates from obtained from mild and severe (hemorrhagic and/or neurologic) from various regions within the endemic area should be sequenced. If possible, human isolates from outside the endemic region should be collected and sequenced as well.
3. **Genetic Near neighbors:**
    a. **Mopeia virus:** The two strains that have been sequenced appear to be diverse (Figure 2A); more strains should be sequenced to allow better a comprehensive view of species diversity.
    b. **Mobala virus:** One strain has been sequenced; more strains should be sequenced to allow better a comprehensive view of species diversity.
    c. **Ippy virus:** One strain has been sequenced. Serological studies indicated that strains are antigenically diverse [17], therefore additional isolates should be sequenced to allow better a comprehensive view of species diversity.

Potential sources of material

Potential sources of human and rodent strains include Mike Bowen, Stuart Nichol, and Pierre Rollin (CDC), Bob Tesh, C.J. Peters and Tom Ksiazek (UTMB), and Manfred Weidman (Institute of Virology in German).

References

1. Briese, T., et al., Genetic detection and characterization of Lujo virus, a new hemorrhagic fever-associated arenavirus from southern Africa. PLoS Pathog, 2009. **5**(5): p. e1000455.
2. Bowen, M.D., et al., *Genetic diversity among Lassa virus strains.* J Virol, 2000. **74**(15): p. 6992-7004.
3. *Centers for Disease Control and Prevention fact sheet, "Lassa Fever,"*. [cited 2009 August 25, 2009]; http://www.cdc.gov/ncidod/dvrd/spb/mnpages/dispages/lassaf.htm.
4. *PathPort, "Pathogen Information: Lassa Virus," Virginia Bioinformatics Institute*. [cited 2009 August 28, 2009]; http://pathport.vbi.vt.edu/pathinfo/pathogens/Lassa_virus_Info.shtml.
5. Fisher-Hoch, S.P. and J.B. McCormick, *Towards a human Lassa fever vaccine.* Rev Med Virol, 2001. **11**(5): p. 331-41.
6. Emmerich, P., S. Gunther, and H. Schmitz, *Strain-specific antibody response to Lassa virus in the local population of west Africa.* J Clin Virol, 2008. **42**(1): p. 40-4.
7. Fichet-Calvet, E., et al., *Fluctuation of abundance and Lassa virus prevalence in Mastomys natalensis in Guinea, West Africa.* Vector Borne Zoonotic Dis, 2007. **7**(2): p. 119-28.
8. Charrel, R.N., X. de Lamballerie, and S. Emonet, *Phylogeny of the genus Arenavirus.* Curr Opin Microbiol, 2008. **11**(4): p. 362-8.
9. Lecompte, E., et al., *Mastomys natalensis and Lassa fever, West Africa.* Emerg Infect Dis, 2006. **12**(12): p. 1971-4.
10. Lecompte, E., et al., *Genetic identification of Kodoko virus, a novel arenavirus of the African pigmy mouse (Mus Nannomys minutoides) in West Africa.* Virology, 2007. **364**(1): p. 178-83.
11. Delgado, S., et al., *Chapare virus, a newly discovered arenavirus isolated from a fatal hemorrhagic fever case in Bolivia.* PLoS Pathog, 2008. **4**(4): p. e1000047.
12. Gunther, S., et al., *Imported lassa fever in Germany: molecular characterization of a new lassa virus strain.* Emerg Infect Dis, 2000. **6**(5): p. 466-76.
13. Carrion, R., Jr., et al., *A ML29 reassortant virus protects guinea pigs against a distantly related Nigerian strain of Lassa virus and can provide sterilizing immunity.* Vaccine, 2007. **25**(20): p. 4093-102.
14. Charrel, R.N., et al., *Phylogeny of New World arenaviruses based on the complete coding sequences of the small genomic segment identified an evolutionary lineage produced by intrasegmental recombination.* Biochem Biophys Res Commun, 2002. **296**(5): p. 1118-24.
15. Vieth, S., et al., *RT-PCR assay for detection of Lassa virus and related Old World arenaviruses targeting the L gene.* Trans R Soc Trop Med Hyg, 2007. **101**(12): p. 1253-64.
16. Emonet, S.F., et al., *Arenavirus genetic diversity and its biological implications.* Infect Genet Evol, 2009. **9**(4): p. 417-29.
17. Meunier, D.Y., et al., *Comparison of Lassa, Mobala, and Ippy virus reactions by immunofluorescence test.* Lancet, 1985. **1**(8433): p. 873-4.

Viral Hemorrhagic Fever: *Marburgvirus*

Summary

Whole genome sequence data is available from human isolates from all of the marburgvirus epidemics and from one species of bat that serves as a marburgvirus vector. Ecological modeling studies and evidence of marburgvirus infection in bats indicate that marburgvirus is present in countries that have not yet reported cases of Marburg hemorrhagic fever [1, 2]; potentially infected samples from this region should be included in the study if possible.

Strains of marburgviruses from the same outbreak may exhibit high overall genetic diversity, therefore multiple isolates should be sequenced from each outbreak. Outbreaks of disease are associated with exposure to bat-infected sites such as caves and mines, and multiple strains of marburgvirus may be present in the thousands of bats that are often present at these sites. In fact, during the 1998 Democratic Republic of Congo (DRC) epidemic, at least nine genetically distinct lineages of virus were introduced into the human population [3]. Two distinct lineages of marburgvirus were detected in a single bat species present at an outbreak site in Uganda [4]. Genetic analysis indicate that genetic diversity is not correlated with geographic distance, therefore multiple samples from the same site and outbreak should be examined for genetic diversity.

Potential sources of this material include CDC, UTMB, Wildlife Conservation Society, and NIAID.

Ecology and epidemiology

Viral hemorrhagic fever outbreaks caused by viruses belonging to the genus *Marburgvirus* (species *Lake Victoria marburgvirus*) have occurred in Angola, DRC, Kenya, Zimbabwe, and Uganda (Figure 1)[5]. Ecological niche modeling shows a broad potential distribution across all arid woodland regions of Africa. Other countries with this ecological niche include Burundi, Ethiopia, Malawi, Mozambique, Rwanda, Tanzania, Zambia, and a small region of N.Cameroon [1].

Most index cases are associated with exposure to bat-infested caves or mines, and serological, genetic, and epidemiological evidence indicates that insectivorous (*Rhinolophus eloquens, Miniopterus inflatus*) and fruit bats (*Rousettus aegyptiacus*) serve as reservoir hosts and vectors. Infected bats have been identified in countries with no history of human marburgvirus cases, as well as in Gabon and Republic of Congo (Figure 2) [2]. Antibodies to marburgvirus have been detected in green vervet monkeys (*Chlorocebus aethiops*) and the first recorded cases of marburgvirus infection were associated with handling infected tissues from green vervet monkeys, however the role of monkeys in marburgvirus transmission is unknown.

The case fatality rate of marburgvirus epidemics ranges from 23-83%. Marburgvirus strains differ in virulence both in terms of case fatality rate and in non human primate animal models, however the genetic basis for virulence difference is unclear [6]. Additionally, the outcome of the infection is influenced by the dose of the virus and route of infection.

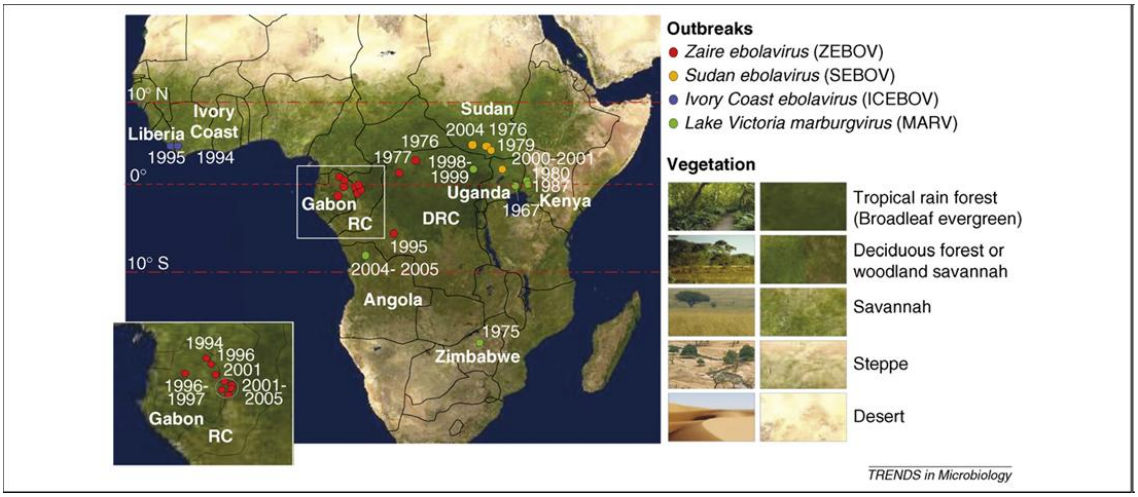


Figure 1. Spatial distribution of human filovirus outbreaks in relation to geographical conditions. The political boundaries and the names of the countries where ebolavirus or marburgvirus outbreaks have been reported are indicated, as are the relevant lines of latitude. The site of each ebolavirus or marburgvirus outbreak is also indicated, as well as the year in which the outbreak occurred. Color patterns in the satellite image are shown at the side, correlated to the ecological conditions they represent [7].

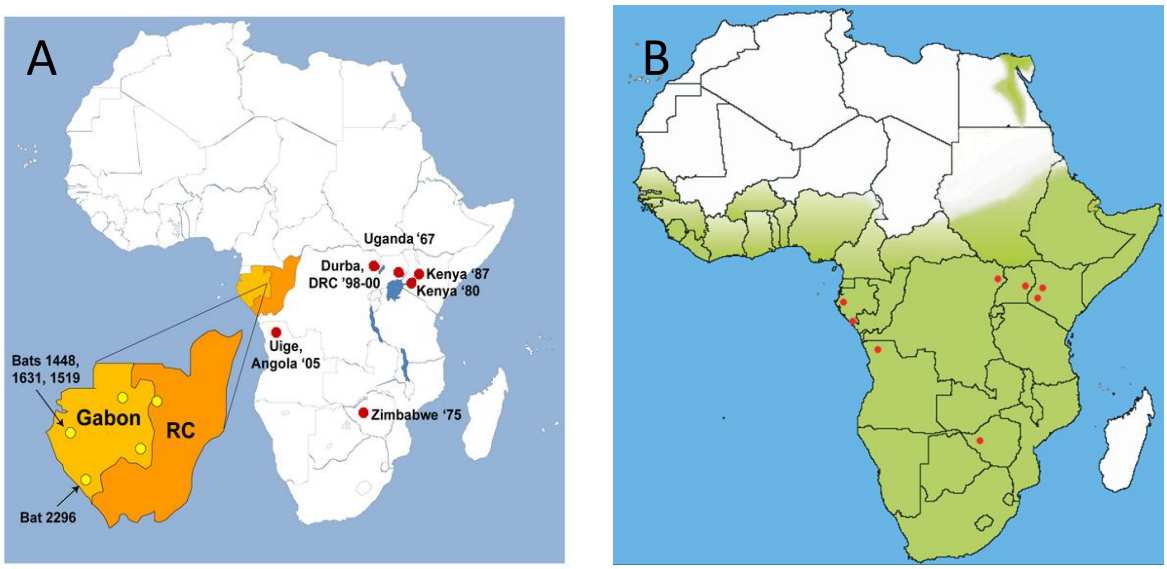


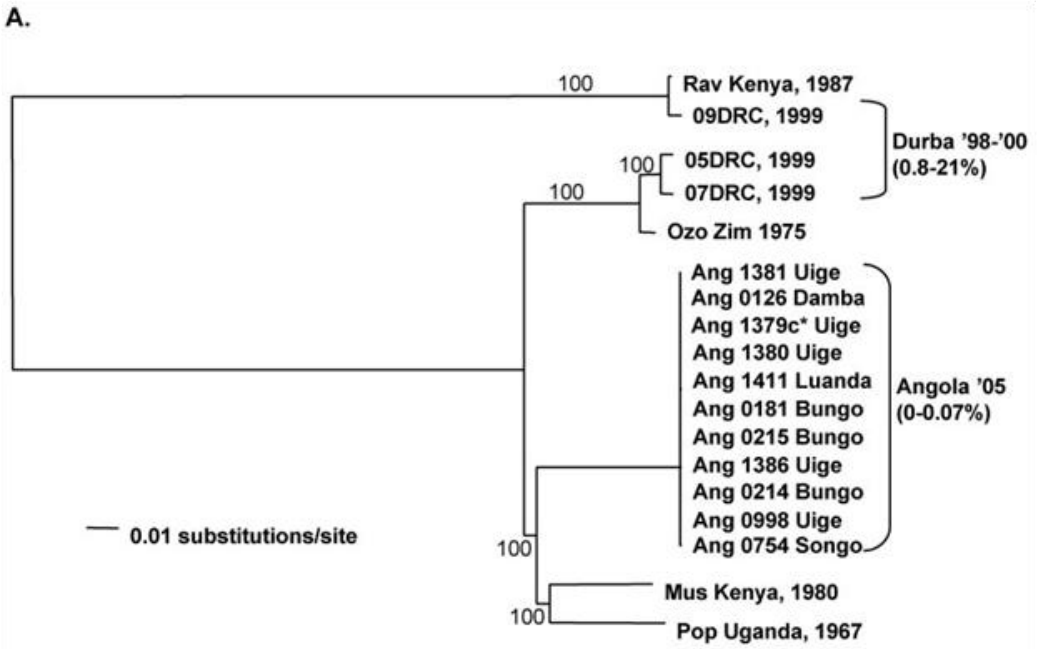
Figure 2. Geographic distribution of marburgvirus outbreaks and marburgvirus bat reservoirs. (A) Animal collection sites in Gabon and Republic of Congo. Animal trap locations in Gabon and Republic of Congo (expansion) are indicated by yellow circles. Also indicated are the locations of the four PCR positive bats (by arrows) and the dates and locations of all known origins of previous marburg virus outbreaks (red circles). (B) Geographic distribution (green shade) of *Rousettus aegyptiacus* in Africa. Red dots indicate the locations of all previous known marburgvirus outbreaks, in addition to the locations of the marburg-positive bats.

Viral taxonomy and genome diversity

Marburgvirus is one of two genera within the family *Filoviridae*; the other genus in this family is *Ebolavirus*. Marburgvirus genomes are ~19,000 nucleotides in length and consist of single-stranded, negative sense RNA. The *Marburgvirus* genus consists of one species: Lake Victoria marburgvirus. Near

neighbors are members of the *Ebolavirus* genus which differs in nucleotide sequence by 55% [6]. Strains of marburgviruses from the same outbreak may exhibit high overall genetic diversity (up to 22%, Figures 3 & 4) [8], however there is only 6.8% nucleotide sequence difference between West Africa (Angola) marburgviruses and East Africa marburgviruses, indicating that the reservoir species in both regions are similar[9].

A.



B.

	Rav	09DRC	05DRC	07DRC	Ozo	Pop	Mus	Ang	EboZ	EboS
Rav	0.00	0.52	21.16	21.22	21.03	21.33	21.09	21.34	68.27	67.88
09DRC		0.00	21.24	21.30	21.11	21.41	21.18	21.41	68.28	67.87
05DRC			0.00	0.80	1.48	7.50	7.22	7.80	67.19	67.42
07DRC				0.00	1.51	7.51	7.27	7.77	67.25	67.51
Ozo					0.00	7.07	6.85	7.47	67.22	67.55
Pop						0.00	5.94	7.14	67.60	67.61
Mus							0.00	6.83	67.43	67.54
Ang								0.00	67.46	67.45
EboZ									0.00	40.47
EboS										0.00

Figure. 3. Genetic diversity of *Marburgvirus*. (A) Maximum-likelihood analysis of full-length genomes with indicated bootstrap values. *, reference specimen. (B) Nucleotide diversity, comparing full-length genome sequences of the indicated filoviruses[9].

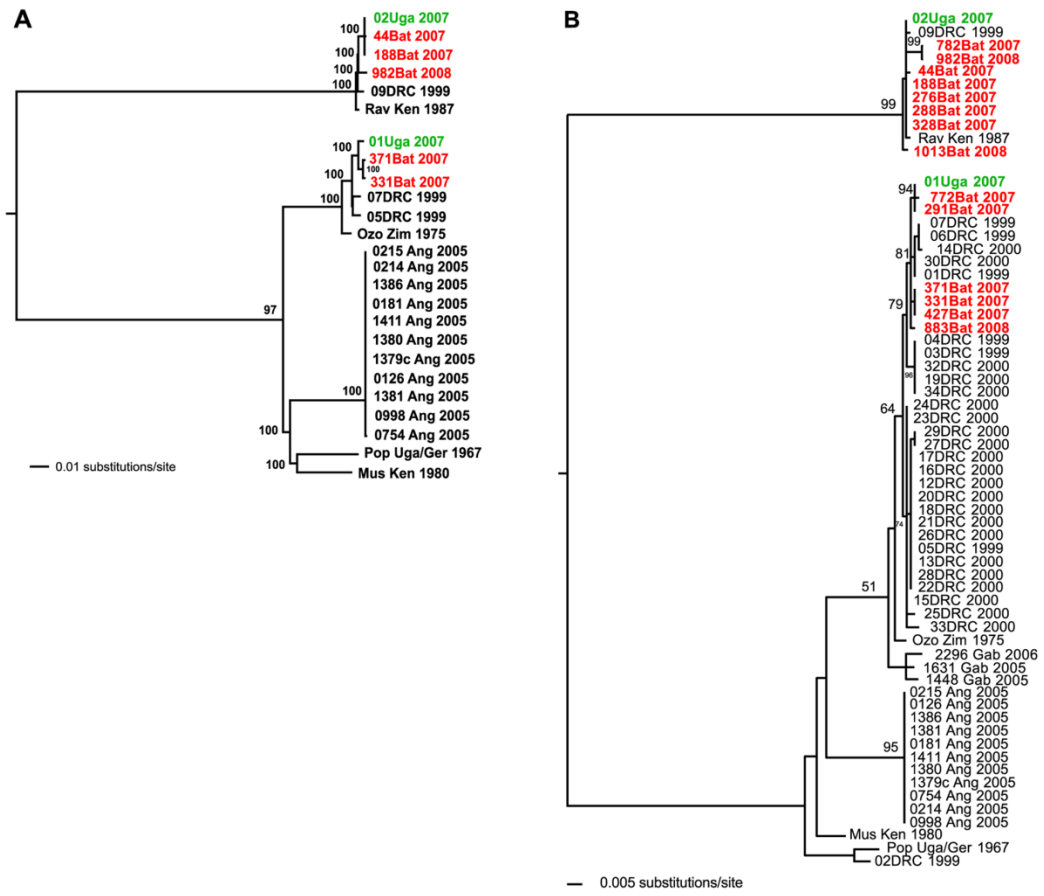


Figure 4. Phylogenetic analysis of full-length or partial genomes of marburg viruses isolated from humans or bats. Marburg virus sequences from 2007 human cases in Uganda are in green, while those from bats are listed in red. (A) Analysis of full-length genomes of five marburg virus bat isolates, 18 historical isolates, and the isolates from patients A and B (01Uga07 and 02Uga07 respectively). (B) Phylogenetic analysis of concatenated NP and VP35 sequence fragments obtained from each bat specimen compared to corresponding regions from 48 historical isolates and those from 01Uga07 and 02Uga07 [4].

Gaps in sequence data

1) Complete genome sequences are available for human samples from all of the marburgvirus outbreaks, as well as from one isolate from one of the three reservoir bat species. Strains of marburgviruses from the same outbreak may exhibit high overall genetic diversity; therefore multiple isolates should be sequenced from each outbreak. Whole genome sequence data is needed for:

i. Additional human isolates from the 1998-2000 DRC epidemic. Isolates should vary geographically and temporally.
ii. Additional isolates from wildlife samples (green vervet monkeys and reservoir bat species) from endemic areas (collected during epidemic and inter-epidemic periods).

2) Ideally, samples should be obtained and analyzed from potentially infected humans, monkey and bat species present in countries that have ecological features that are characteristic of marburgvirus endemic regions but have yet to report cases. For example, a recent study by Tong et al. (2009)

describes the collection of tissue samples from bats in Kenya for the purposes of identifying new coronaviruses [8]. The bat species collected include all three bat species suspected of vectoring marburgvirus; any available samples from this study should be screened for infection with filoviruses.

Potential sources of material

1) CDC (Stuart Nichol, Anthony Sanchez, Jonathan Towner): Human and animal samples (E. Leroy collection)
2) UTMB (Ksiazek): Human and animal samples (E. Leroy collection)
3) NIAID at Rocky Mountain Labs (Heinz Feldmann): Human and animal samples
4) Sofi Ibrahim and Lisa Hensley at USAMRIID
5) Kate Rubins at Whitehead Institute
6) Manfred Weidman at the Institute of Virology in Germany

References

1. Peterson, A.T., et al., *Geographic potential for outbreaks of Marburg hemorrhagic fever.* Am J Trop Med Hyg, 2006. 75(1): p. 9-15.
2. Towner, J.S., et al., *Marburg virus infection detected in a common African bat.* PLoS One, 2007. 2(1): p. e764.
3. Bausch, D.G., et al., *Marburg hemorrhagic fever associated with multiple genetic lineages of virus.* N Engl J Med, 2006. 355(9): p. 909-19.
4. Towner, J.S., et al., *Isolation of genetically diverse marburg viruses from egyptian fruit bats.* PLoS Pathog, 2009. 5(7): p. e1000536.
5. Feldmann, H., et al., *Ebola virus ecology: a continuing mystery.* Trends Microbiol, 2004. 12(10): p. 433-7.
6. Sanchez, A., Geisbert, T.W., Feldmann H., *Filoviridae: Marburg and Ebola Viruses*, in *Fields Virology*, D.M. Knipe, Howley, P.M. , Editor. 2007, Lippincott Williams & Wilkins: Philadelphia. p. 1409-1448.
7. Groseth, A., H. Feldmann, and J.E. Strong, *The ecology of Ebola virus.* Trends Microbiol, 2007. 15(9): p. 408-16.
8. Tong, S., et al., *Detection of novel SARS-like and other coronaviruses in bats from Kenya.* Emerg Infect Dis, 2009. 15(3): p. 482-5.
9. Towner, J.S., et al., *Marburgvirus genomics and association with a large hemorrhagic fever outbreak in Angola.* J Virol, 2006. 80(13): p. 6497-516.

Viral Hemorrhagic Fever: *Ebolavirus*

Summary

Whole genome sequence data is needed from human and animal hosts and vectors. Currently, there is no whole genome sequence data available for 8 of the 18 ebolavirus outbreaks or from infected (viral RNA positive) bat and ape samples. Ecological studies suggest that ebolavirus is present in countries that have not yet reported cases of ebolavirus hemorrhagic fever (such as Cameroon and Ghana); potentially infected samples from this region should be included in the study if possible. Potential sources of samples include CDC, UTMB, Wildlife Conservation Society, and NIAID.

Ecology and epidemiology

Viral hemorrhagic fever outbreaks caused by viruses in the genus *Ebolavirus* occur regularly in the tropical regions of Central Africa (Figure 1) [1]. Ebolavirus infection of non-human primates such as gorillas and chimpanzees may cause fatal disease and human disease outbreaks may be preceded by detection of the carcasses of diseased apes and duikers (forest antelope) which also may succumb to the infection. Indeed, the high fatality rate of ebolavirus infection in apes threatens to wipe out entire gorilla and chimpanzee populations present in ebolavirus endemic regions. Handling of infected animal tissues may result in spread of the disease to humans.

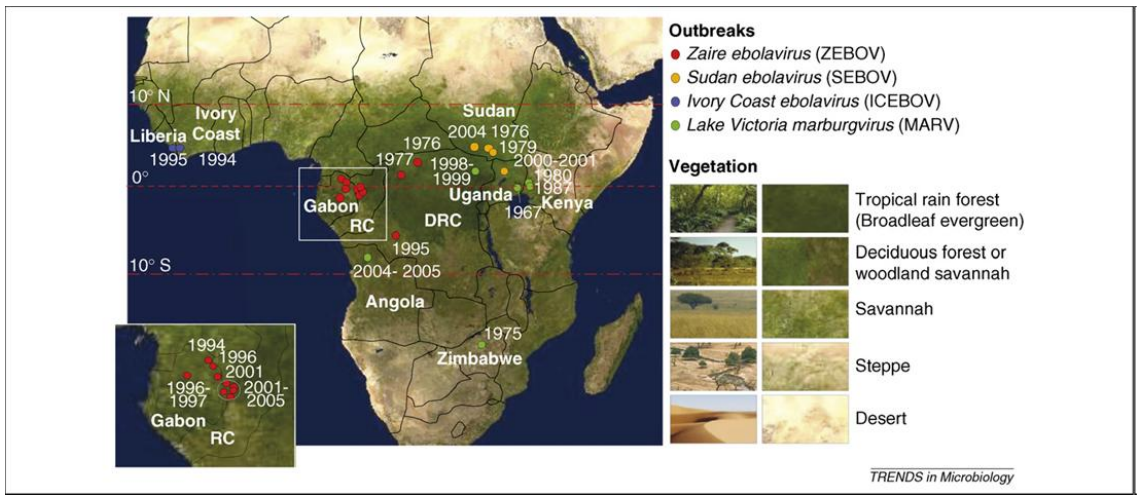


Figure 1. Spatial distribution of human filovirus outbreaks in relation to geographical conditions. The political boundaries and the names of the countries where ebolavirus or marburgvirus outbreaks have been reported are indicated, as are the relevant lines of latitude. The site of each ebolavirus or marburgvirus outbreak is also indicated, as well as the year in which the outbreak occurred. Color patterns in the satellite image are shown at the side, correlated to the ecological conditions they represent [2].

Insectivorous and fruit bats are indicated as the reservoir species based on genetic, epidemiological, and serological evidence [3]. The geographical range of the bat reservoir species overlaps that of epidemics but also includes tropical regions of Central Africa that have no reported ebolavirus cases (Figure 2). Infected bats are detected more frequently during outbreaks periods but are also present in between outbreaks. Additionally, animals with antibodies to ebolavirus have been identified in countries where no human cases have occurred, for example, there are seropositive chimpanzees in Cameroon but no reported ebolavirus cases.

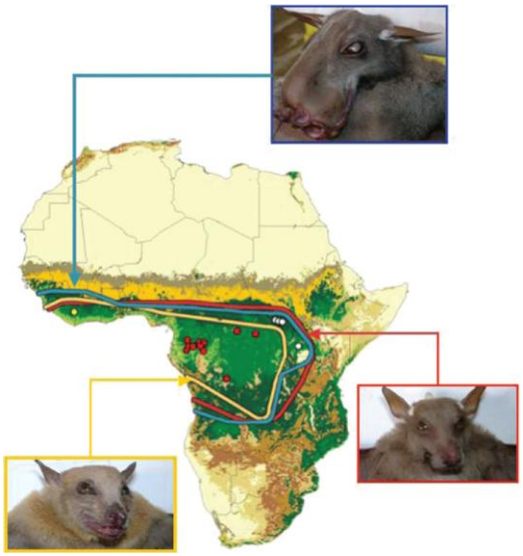


Figure 2. Distribution of the three bat species that are potential reservoirs of ebolavirus [4]. Geographic distribution (inside colored lines) of the fruit bats *Hypsignathus monstrosus* (blue), *Epomops franqueti* (red) and *Myonycteris torquata* (yellow).

Reston ebolavirus is endemic in the Philippines and has been transported to quarantine facilities in the United States and Italy through the importation of infected non-human primates. The outbreaks of disease caused by importation of this virus were extremely lethal for the infected monkeys but did not cause disease in the humans exposed to the virus. Recently *Reston ebolavirus* was detected in Philippine domestic swine that were co-infected with porcine reproductive and respiratory syndrome virus and were displaying severe respiratory signs [5]. The role of swine in the natural transmission cycle of *Reston ebolavirus* is unclear.

Viral taxonomy and genome diversity

Ebolavirus is one of two genera within the family *Filoviridae*; the other genus in this family is *Marburgvirus*. Ebolavirus genomes are ~19,000 nucleotides in length and consist of single-stranded, negative sense RNA. The *Ebolavirus* genus consists of the following species: *Sudan ebolavirus, Zaire ebolavirus, Ivory Coast ebolavirus, Reston ebolavirus,* and *Bundibugyo ebolavirus. Reston ebolavirus* is the only *Ebolavirus* species that is not associated with human disease and this nonpathogenic species differs by 30-40% as compared to the four species of *Ebolavirus* that are human pathogens [6]. Sequence variation within the *Reston ebolavirus* strains with complete genomes sequenced shows that the strains are very similar with about 96% nucleotide sequence identity (Figure 3) [6]. Near neighbors are members of the *Marburgvirus* genus which differ in nucleotide sequence by 55% [7].

A

Nucleotide Similarity	Reston08-A	Reston08-C	Reston08-E	Reston96
Reston08-A				
Reston08-C	95.5%			
Reston08-E	96.9%	95.8%		
Reston96	97.7%	96.5%	98.70%	
Reston89	97.7%	96.9%	97.90%	98.80%

Amino acid changes	Reston08-A	Reston08-C	Reston08-E	Reston96
Reston08-A 5209 aa				
Reston08-C 5201 aa	193			
Reston08-E 5201 aa	122	193		
Reston96 5201 aa	98	171	64	
Reston89 5201 aa	105	148	92	61

B

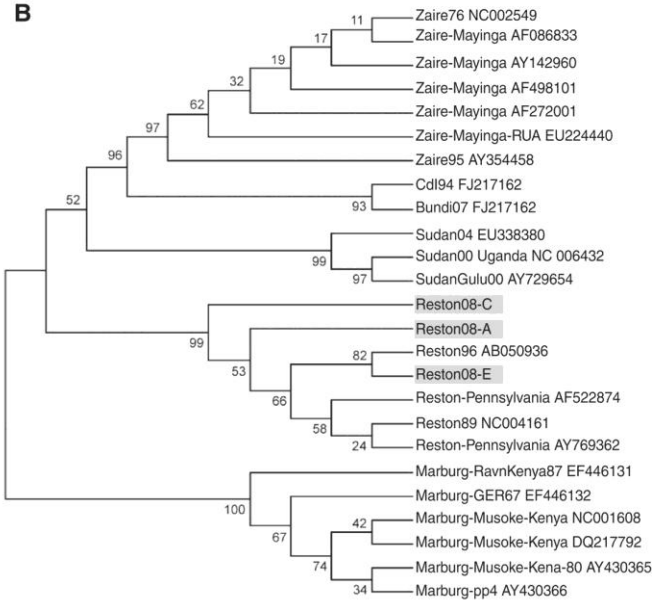


Figure 3. Phylogeny of *Filoviridae* with emphasis on the newly discoved *Reston ebolavirus* strains isolated from swine (Reston08-A, Reston08-C, and Reston08-E). (A) Full-length genomic sequences with the exception of the defined 5' and 3' termini. (B) A consensus neighbor-joining tree drawn without distance topology illustrates the independent branching of the three 2008 Philippine swine viruses within the *Reston ebolavirus* clade [5].

Although the ebolavirus genome has high genetic stability and is preserved temporally and spatially within a species, species of *Ebolavirus* differ genetically from each other by up 37-41%. The mortality rate also differs between species with between 57-90% mortality rates for *Zaire ebolavirus* epidemics and 41-65% mortality rates for *Sudan ebolavirus* epidemics [7]. Interestingly, the genotypes of viruses obtained from survivors, fatal cases, and asymptomatic cases do not appear to differ; the GP and NP gene sequences are identical for fatalities, survivors and asymptomatically infected individuals during 1996-97 Gabon outbreak [8]. Distinct strains may circulate during an outbreak indicating multiple introductions, for example, outbreaks that occurred on the Gabon-Republic of Congo border from 2001-2003 were caused by many genetically different strains.

Gaps in sequence data

1) Complete genome sequences are available for 8 of the 18 ebolavirus hemorrhagic fever outbreaks. Whole genome sequence data is needed for:

i. Gabon-Republic of Congo outbreaks (n=5), DRC 2007, and Gabon 1994, 1996-97

2) All genomes sequenced so far are from humans; there is no complete sequence data available for viruses present in wildlife hosts or vectors:

i. Wildlife samples (ape, bat, etc.) from endemic areas (collected during epidemic and inter-epidemic periods) should be screen for the presence of filovirus genomes.

3) Ideally, samples should be obtained and analyzed from potentially infected humans, apes and bat species present in countries that have ecological features that are characteristic of ebolavirus endemic regions but have yet to report cases.

Potential sources of material

1) CDC (Stuart Nichol, Anthony Sanchez, Jonathan Towner): Human and animal samples (E. Leroy collection)
2) UTMB (Ksiazek): Human and animal samples (E. Leroy collection)
3) NIAID at Rocky Mountain Labs (Heinz Feldmann): Human and animal samples
4) Sofi Ibrahim and Lisa Hensley at USAMRIID
5) Kate Rubins at Whitehead Institute
6) Manfred Weidman at the Institute of Virology in Germany

References

1. Feldmann, H., et al., *Ebola virus ecology: a continuing mystery.* Trends Microbiol, 2004. 12(10): p. 433-7.
2. Groseth, A., H. Feldmann, and J.E. Strong, *The ecology of Ebola virus.* Trends Microbiol, 2007. 15(9): p. 408-16.
3. Gonzalez, J.P., X. Pourrut, and E. Leroy, *Ebolavirus and other filoviruses.* Curr Top Microbiol Immunol, 2007. 315: p. 363-87.
4. Leroy, E.M., et al., *Fruit bats as reservoirs of Ebola virus.* Nature, 2005. 438(7068): p. 575-6.
5. Barrette, R.W., et al., *Discovery of swine as a host for the Reston ebolavirus.* Science, 2009. 325(5937): p. 204-6.
6. Groseth, A., et al., *Molecular characterization of an isolate from the 1989/90 epizootic of Ebola virus Reston among macaques imported into the United States.* Virus Res, 2002. 87(2): p. 155-63.
7. Sanchez, A., Geisbert, T.W., Feldmann H., *Filoviridae: Marburg and Ebola Viruses*, in *Fields Virology*, D.M. Knipe, Howley, P.M. , Editor. 2007, Lippincott Williams & Wilkins: Philadelphia. p. 1409-1448.
8. Leroy, E.M., et al., *Sequence analysis of the GP, NP, VP40 and VP24 genes of Ebola virus isolated from deceased, surviving and asymptomatically infected individuals during the 1996 outbreak in Gabon: comparative studies and phylogenetic characterization.* J Gen Virol, 2002. 83(Pt 1): p. 67-73.

Junin virus: Argentine hemorrhagic fever

Summary

A large number of Junin virus strains have been isolated from humans and rodents, however only four strains have been fully sequenced, all of which are laboratory passaged strains obtained from humans decades ago. Additionally, three of the four strains sequenced are from derived from the same isolate. Analysis of partial genome sequence data of 39 strains of Junin virus obtained from human and rodent hosts indicates that there is up to 13% nucleotide differences between strains [1]. Whole genome analysis is needed for viral strains isolated more recently from humans and from rodent vectors from different geographical regions of the endemic area. Isolates from both mild and severe cases of Argentine hemorrhagic fever (AHF) should be sequenced as should strains of near neighbors, Machupo virus and Tacaribe virus.

Ecology and epidemiology

AHF is endemic in the agriculturally-developed pampas region of Argentina and is vectored primarily by field mice (*Calomys musculinus)* which are persistently infected and shed virus in their excreta. Junin virus causes a chronic infection in *Calomys musculinus* and these rodents serve as the primary reservoir [2]. Other local rodent species such as *Calomys laucha* and *Bolomys obscuris* are also known to become infected.

AHV is a seasonal disease associated with the grain harvest during which agricultural workers are most likely to encounter the infected field mice. There has been a progressive expansion of the geographic range of AHF outbreaks since its discovery in 1958 (Figure 1. A). Outbreaks now occur east, northeast and northwest of the original outbreak, however the incidence of outbreaks has decreased. Agricultural workers have the highest risk for disease and between 300-1000 cases of AHF are diagnosed each year [3]. Use of a vaccine has greatly decreased the disease incidence in at-risk agricultural workers[4].

The range of the principal rodent vector of Junin virus extends beyond the AHF endemic region and infected rodents are occasionally detected outside the endemic region. Genetic analysis of *Calomys musculinus* populations indicate that Junin virus may become extinct in populations of mice and reintroduced periodically via horizontal transmission mechanisms such as biting [5]. Outbreaks of AHF may occur due to increased rodent densities due to agricultural development and favorable weather conditions [6].

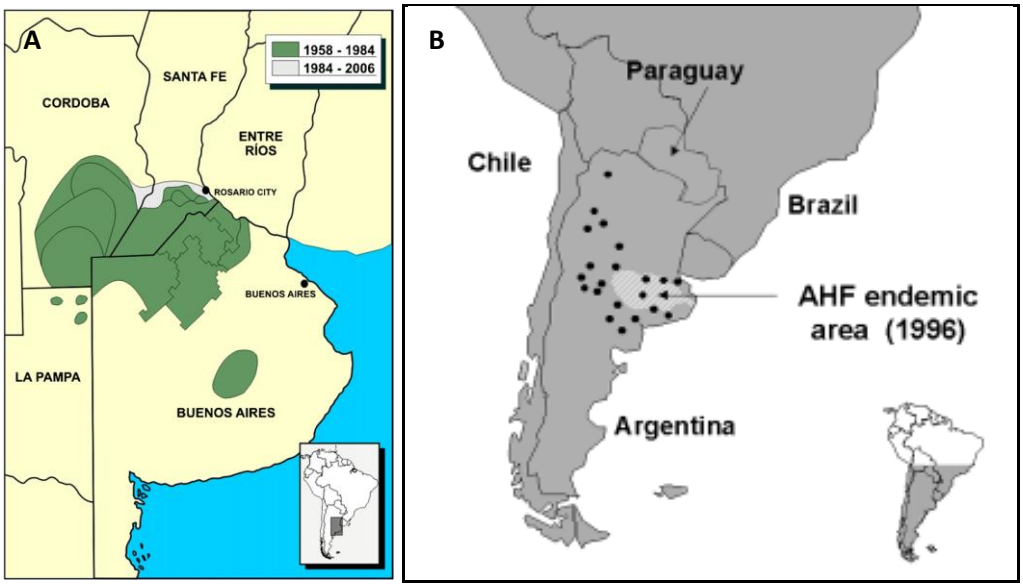


Figure 1. The endemic area of AHF in relationship to (A) recent expansion of the region of endemicity, and (B) the distribution of the rodent vector. (B) Distribution of *Calomys musculinus* (dots), and disease-endemic area of AHF (shaded) [4, 7].

Viral taxonomy and genome diversity

Junin virus is a member of the *Arenaviridae* family, which includes two serocomplexes. The Old World serocomplex is composed of 5 species of viruses, including Lassa virus, and the New World serocomplex contains 16 virus species, including Machupo virus, Junin virus, and Guanarito virus (Figure 2). The arenavirus genome consists of two single-stranded segments of RNA, the S segment (3.4 kb) and the L segment (7.2 kb). Both segments are arranged in an ambisense orientation.

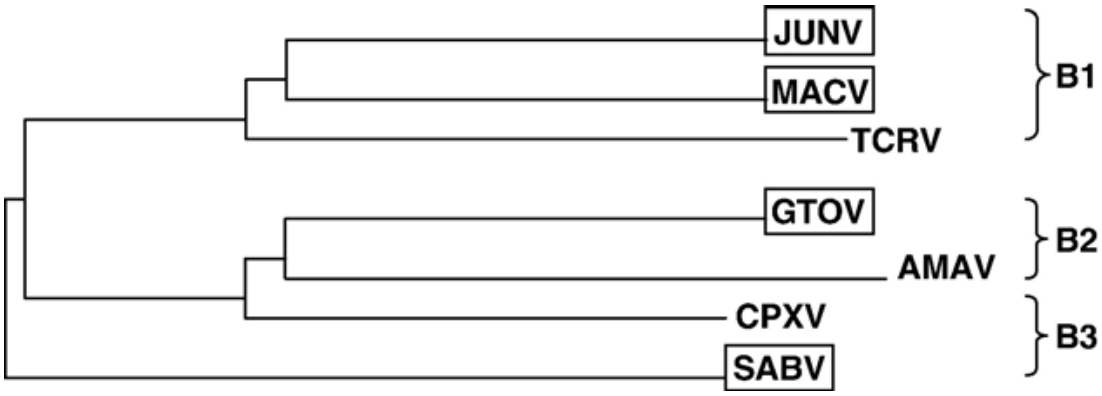


Figure 2. Phylogeny of New World arenaviruses (Clade B). The arenaviruses are classified into two major groups, the Old World and the New World/Tacaribe complex viruses. The New World viruses are further subdivided into three Clades, with the Clade B lineage of particular interest since it contains four significant human pathogens: Junin (JUNV), Machupo (MACV), Guanarito (GTOV) and Sabia (SABV) viruses. The strains that are pathogenic for humans (boxed) are distributed among the three sub-lineages (phylogenetic analysis of amino acid sequences in the glycoprotein gene)[8].

Analysis of partial genome sequence data of 39 strains of Junin virus isolates obtained from human and rodent hosts indicates that there is up to 13% nucleotide differences between strains (Figure 3) [1]. Near neighbors of Junin virus include Machupo virus and Tacaribe virus (Figure 2, Table 1). Machupo virus is the cause of Bolivian Hemorrhagic fever and differs from Junin virus in nucleotide sequence by 25-31% [9]. Tacaribe virus is a nonpathogenic arenavirus vectored by bats differs from Junin virus in nucleotide sequence by 31-33% [9]. Whole genome sequence data is available for one Tacaribe virus isolate.

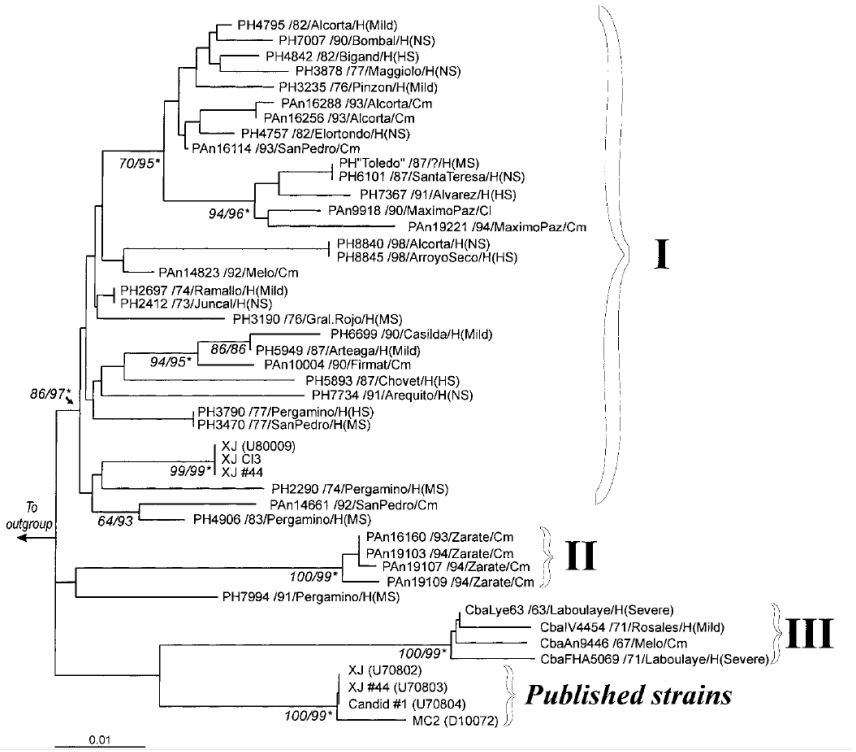


Figure 3. Phylogenetic relationships among Junin virus strains. The phylogenetic tree, generated using nucleotide sequence differences of the 511-nt fragment of the NP gene, revealed three distinct clades. The first clade was composed of strains from the center of the endemic area; the second clade contained 4 strains isolated from 1963 to 1971 from the western-most edge of the endemic area; and the third clade contained 4 strains from *Calomys musculinus* trapped in the northeastern edge of the endemic area. Junin sequences from XJ, XJ #44, and Candid #1 strains formed a separate clade [1]. Strains are described as follows: strain designation/year/locality of origin/source (human or rodent)/(clinical form of AHF [for humans]); H, human; Cm, *Calomys musculinus*; Cl, *Calomys laucha*; HS, hemorrhagic severe; NS, neurological severe; MS, mixed severe.

Virus	PIRV	PICV	GTOV	JUNV	MACV	SABV	TCRV	LASV	LCMV
Pirital	–	76,8	44,1	42,4	43,5	45,3	43,0	46,6	41,6
Pichindé	61,7	–	41,9	39,1	40,2	41,1	39,8	42,0	40,4
Guanarito	44,3	44,3	–	64,9	62,8	60,0	67,4	38,4	39,1
Junin	45,7	43,4	59,4	–	76,6	52,1	70,2	41,9	40,2
Machupo	45,5	43,7	60,6	73,4	–	56,4	73,4	41,9	39,1
Sabiá	44,2	43,7	59,3	60,5	61,3	–	54,7	37,5	40,4
Tacaribe	44,6	43,9	58,7	70,2	71,0	60,8	–	39,5	34,5
Lassa	36,1	36,3	35,4	35,5	35,5	35,7	35,1	–	53,9
Lymphocytic choriomenin-gitis	37,4	34,9	35,0	35,1	34,8	33,9	34,4	47,5	–

Table 1. Amino acid sequence identities among the Z proteins (% amino acid sequence identity) and among the RNA-dependent RNA polymerases (% amino acid sequence identity) of nine arenaviruses [10]. GTOV, Guanarito virus; JUNV, Junín virus; MACV, Machupo virus; PICV, Pichindé virus; PIRV, Pirital virus; SABV, Sabiá virus; TCRV, Tacaribe virus.

Gaps in sequence data

A large number of strains have been isolated from both humans and rodents, however few strains have been fully sequenced, all of which are laboratory passaged strains obtained from human decades ago. Both hemorrhagic and neurologic symptoms may be associated with human Junin virus infections, and strains differ in pathogenicity in animal models as well as the human host [11], however the genetic determinants of virulence have yet to be identified. The sequenced strains are the Romero strain, a

human isolate from a severe but nonlethal AHF case displaying both hemorrhagic and neurological symptoms (isolated 1976 or 1977); the XJ strain which was isolated from a severe 1958 AHF case with hemorrhagic symptoms and passaged 37 times in suckling mice before sequence analysis [12]; and strains XJ13 and Candid#1 which are attenuated strains that were derived from the XJ strain by further laboratory passage [13].

Suggested strains for whole genome sequence analysis are:

1. **Rodent vectors**: Recent and historical isolates from the principal host, *Calomys musculinus* should be sequenced. The following strains are genetically diverse as indicated by sequence analysis of partial genome which defined phylogenetic clade (I-III), location (northwest (NW), north central (NC), northeast (NE), and year [1]:

a. PAn19221 (Clade I, NC, 1994)
b. Pan16114 (Clade I, NE, 1993)
c. Pan10004 (Clade I, NC, 1990)
d. PAn19107 (Clade II, NE, 1994)
e. Cba An9446 (Clade III, NW, 1967)

2. **Human hosts**: More recent isolates from obtained from mild and severe (hemorrhagic and/or neurologic) from various regions within the endemic area should be sequenced. The following strains fit this criteria and are shown to be genetically diverse samples from sequence analysis of GP1 gene [1]:

a. Strain PH7007, 1990, neurological symptoms, north central region
b. Strain PH8845, 1998, hemorrhagic symptoms, northeast/central region
c. Strain PH6699, 1991, mild symptoms, north central region

3. Genetic Near neighbors:

a. Tacaribe virus (only one complete sequence available)
b. Machupo virus (see Machupo report section)

Potential sources of material

Potential sources of strains include Mike Bowen and Stuart Nichol, (CDC); Steve St. Jeor (Univ. of Nevada, Reno); and Bob Tesh and Tom Ksiazek (UTMB).

References

1. Garcia, J.B., et al., Genetic diversity of the Junin virus in Argentina: geographic and temporal patterns. Virology, 2000. **272**(1): p. 127-36.
2. Salazar-Bravo, J., et al., Natural nidality in Bolivian hemorrhagic fever and the systematics of the reservoir species. Infect Genet Evol, 2002. **1**(3): p. 191-9.
3. Enria, D.A., A.M. Briggiler, and Z. Sanchez, Treatment of Argentine hemorrhagic fever. Antiviral Res, 2008. **78**(1): p. 132-9.
4. Enria, D.A., Briggiler, A.M., Feuilade, M.R., An overview of the epidemiological, ecological and preventive hallmarks of Argentine haemorrhagic fever (Junin virus). Bulletin de l'Institut Pasteur, 1998. **96**: p. 103-114.

5. Gonzalez Ittig, R.E. and C.N. Gardenal, Recent range expansion and low levels of contemporary gene flow in Calomys musculinus: its relationship with the emergence and spread of Argentine haemorrhagic fever. Heredity, 2004. **93**(6): p. 535-41.
6. Mills, J.N., et al., A longitudinal study of Junin virus activity in the rodent reservoir of Argentine hemorrhagic fever. Am J Trop Med Hyg, 1992. **47**(6): p. 749-63.
7. Mills, J.N. and J.E. Childs, Ecologic studies of rodent reservoirs: their relevance for human health. Emerg Infect Dis, 1998. **4**(4): p. 529-37.
8. Oldenburg, J., et al., Differences in tropism and pH dependence for glycoproteins from the Clade B1 arenaviruses: implications for receptor usage and pathogenicity. Virology, 2007. **364**(1): p. 132-9.
9. Delgado, S., et al., Chapare virus, a newly discovered arenavirus isolated from a fatal hemorrhagic fever case in Bolivia. PLoS Pathog, 2008. **4**(4): p. e1000047.
10. Cajimat, M.N., et al., Genetic diversity among Bolivian arenaviruses. Virus Res, 2009. **140**(1-2): p. 24-31.
11. Kenyon, R.H., et al., Heterogeneity of Junin virus strains. Med Microbiol Immunol, 1986. **175**(2-3): p. 169-72.
12. Castilla, V., M. Contigiani, and S.E. Mersich, Inhibition of cell fusion in Junin virus-infected cells by sera from Argentine hemorrhagic fever patients. J Clin Virol, 2005. **32**(4): p. 286-8.
13. Goni, S.E., et al., Genomic features of attenuated Junin virus vaccine strain candidate. Virus Genes, 2006. **32**(1): p. 37-41.

Bolivian hemorrhagic fever: Machupo virus

Summary

Whole genome sequence data exists for two strains of Machupo virus. Both sequenced strains are from human cases and were isolated decades ago (1963 and 1971). Phylogenetic analysis of the N protein gene sequence data from human and rodent strains obtained between 1963-2000 identified eight distinct genetic lineages within the Machupo species with up to 13% nucleotide difference between strains [1]. Suggested strains for whole genome sequence analysis include isolates obtained from the rodent vector (vesper mouse) as well as isolated obtained from more recent human cases.

Ecology and epidemiology

Machupo virus is the causative agent of Bolivian hemorrhagic fever (BHF). BHF is endemic in the savannas of northeast Bolivia (Figure 1). Machupo virus is vectored by vesper mice (*Callomys callosus*) and is typically transmitted to humans through inhalation of virus-infected rodent excrement. Vesper mice prefer living in open areas such as land plots cleared for crops and people are exposed to the mice during crop harvest. Apparently only a geographical subgroup (cryptic species) within the *Callomys callosus* species are known to transmit Machupo virus. Because only a subgroup of vesper mice serve as vectors of Machupo virus, BHF is not endemic throughout the entire range of vesper mice [2].

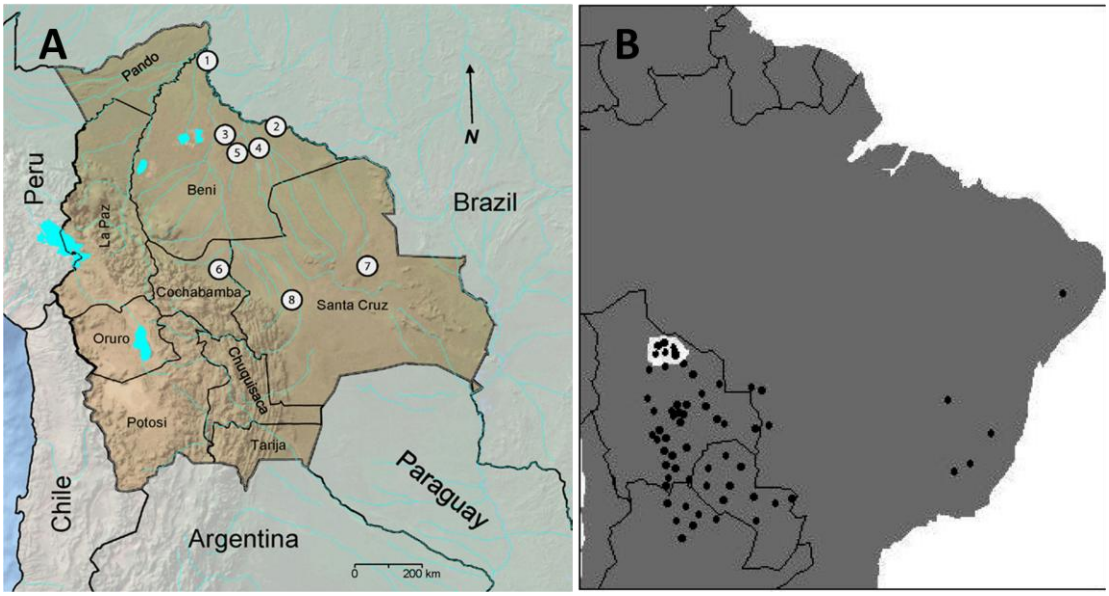


Figure 1. Geographical range of (A) human Machupo virus infections as compared to other Bolivian arenaviruses, and (B) the greographical range of the rodent vector of Machupo virus. (A) Map showing the 9 departments in Bolivia and the locations of (1) Guayaramerín, (2) Magdalena, (3) San Joaquín, (4) Huacaraje, (5) SanRamón,(6) Chapare, (7) San Ignacio, and (8) Juan Latino. Strains of Machupo virus from hemorrhagic fever cases were sampled from Guayaramerín, Huacaraje, Magdalena, San Joaquín, San Ramón, and Villamontes (location on the map not known) and from vesper mice captured at sites near San Ramón. Two other species of arenavirus are present in Bolivia, Latino virus, which was isolated near Juan Latino and San Ignacio, and Chapare virus, a newly discovered hemorrhagic fever virus which was isolated near the Chapare river [1]. (B) The distribution of *C. callosus*. Black dots represent localities based on museum specimens. The white area is the region where BHF has been reported in the Beni Department of Bolivia [2].

Sporadic cases are reported annually from the endemic region (Beni department of Bolivia). For example, a search of the ProMED archives shows 10 cases reported in 1994, 3 cases in 1996, 8 cases in 1999, 6 cases in 2000, 2 cases in 2004, and 5 cases 2007. An outbreak occurred in San Joaquin, Bolivia during 1963-1964 with 637 cases and an 18% case fatality rate.

Viral taxonomy and genome diversity

Machupo virus is a member of the *Arenaviridae* family, which includes two serocomplexes. The Old World serocomplex is composed of 5 species of viruses, including Lassa virus, and the New World serocomplex contains 16 virus species, including Machupo virus, Junin virus, and Guanarito virus. Arenaviruses have a genome of ~ 10.6 kilobases that is composed of two segments of RNA with genes coded an ambisense arrangement. Phylogenetic analysis of the N protein gene sequence resulted in the identification of eight distinct genetic lineages within the Machupo species with up to 13% nucleotide difference between strains [1].

Near neighbors of Machupo virus include Junin virus and Tacaribe virus (Figure 2, Table 1). Junin virus is the cause of Argentine hemorrhagic fever and differs from Machupo virus by 25-31%.[3] As stated in the Junin virus section of this report, whole genome sequence data for Junin virus strains is limited with 3 of

the 4 strains sequenced derived from the same isolate, and all the sequences obtained from human samples collected decades ago. Tacaribe virus is a nonpathogenic arenavirus vectored by bats, and has 31-33% nucleotide difference with Machupo virus. Whole genome sequence data is available for one Tacaribe virus isolate.

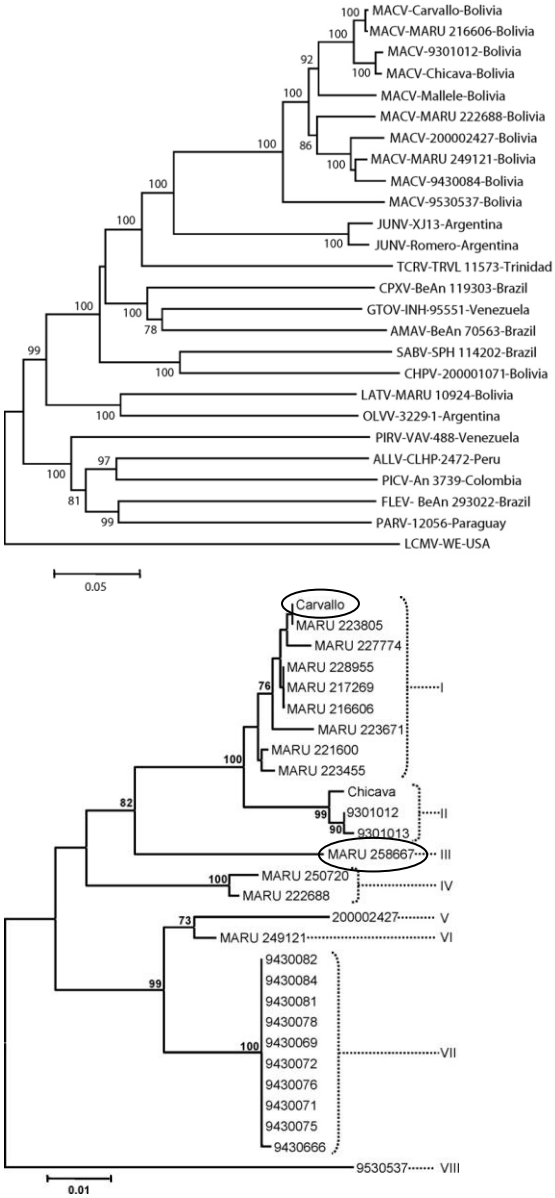


Figure 2. Phylogenetic relationships among 10 strains of Machupo virus (upper dendrogram) and 15 other New World arenaviruses based on analysis of the full-length nucleocapsid protein gene sequences, and 28 strains of Machupo virus based on nucleocapsid protein gene sequences (lower dendrogram). The Roman numerals indicate the 8 major phylogenetic lineages represented by the 28 strains [1]. Machupo virus isolates that have whole genome sequence available are circled. ALLV, Allpahuayo virus; AMAV, Amapari virus; CHPV, Chapare virus; CPXV, Cupixi virus; FLEV, Flexal virus; GTOV, Guanarito virus; JUNV, Junín virus; LATV, Latino virus; MACV, Machupo virus; OLVV, Oliveros virus; PARV, Paraná virus; PICV, Pichindé virus; PIRV, Pirital virus; SABV, Sabiá virus; TCRV, Tacaribe virus.

Virus	PIRV	PICV	GTOV	JUNV	MACV	SABV	TCRV	LASV	LCMV
Pirital	–	76.8	44.1	42.4	43.5	45.3	43.0	46.6	41.6
Pichindé	61.7	–	41.9	39.1	40.2	41.1	39.8	42.0	40.4
Guanarito	44.3	44.3	–	64.9	62.8	60.0	67.4	38.4	39.1
Junin	45.7	43.4	59.4	–	76.6	52.1	70.2	41.9	40.2
Machupo	45.5	43.7	60.6	73.4	–	56.4	73.4	41.9	39.1
Sabiá	44.2	43.7	59.3	60.5	61.3	–	54.7	37.5	40.4
Tacaribe	44.6	43.9	58.7	70.2	71.0	60.8	–	39.5	34.5
Lassa	36.1	36.3	35.4	35.5	35.5	35.7	35.1	–	53.9
Lymphocytic choriomeningitis	37.4	34.9	35.0	35.1	34.8	33.9	34.4	47.5	–

Table 1. Amino acid sequence identities among the Z proteins (% amino acid sequence identity) and among the RNA-dependent RNA polymerases (% amino acid sequence identity) of nine arenaviruses.[4]

Gaps in sequence data

Complete genome sequences are available from two cases of hemorrhagic fever which occurred in 1963 and 1971. Whole genome sequence data is needed for:

i) Isolates from more recent human cases
ii) Isolates from vesper mice

Potential sources of material

1) CDC (Nichol): Human and rodent samples
2) UTMB (Fulhorst, Ksiazek, Weaver, Tesh, Peters): Human and rodent samples
3) Univ. of New Mexico (Salazar-Bravo, Yates): Rodent samples

References

1. Cajimat, M.N., et al., *Genetic diversity among Bolivian arenaviruses.* Virus Res, 2009. **140**(1-2): p. 24-31.
2. Salazar-Bravo, J., et al., *Natural nidality in Bolivian hemorrhagic fever and the systematics of the reservoir species.* Infect Genet Evol, 2002. **1**(3): p. 191-9.
3. Delgado, S., et al., *Chapare virus, a newly discovered arenavirus isolated from a fatal hemorrhagic fever case in Bolivia.* PLoS Pathog, 2008. **4**(4): p. e1000047.
4. Cajimat, M.N. and C.F. Fulhorst, *Phylogeny of the Venezuelan arenaviruses.* Virus Res, 2004. **102**(2): p. 199-206.

Guanarito virus: Venezuelan hemorrhagic fever

Summary

The only fully sequenced genome of Guanarito virus is an isolate from a human case of Venezuelan hemorrhagic fever (VHF) that occurred in 1990, therefore more recent isolates from humans and rodent vectors need to be sequenced. Isolates of near neighbors Amapari, Cupixi, Tacaribe, and Sabia viruses should be sequenced as one completely sequenced genome exists for each of these viruses. Sources of viral isolates include CDC (Bowen, Nichols) and UTMB (Fulhorst, Weaver, Tesh), however, in the case of near neighbor species, it is likely that no additional virus isolates are available.

Ecology and epidemiology

Guanarito virus was first discovered in 1989 in a western region of Venezuela in which tropical forest was cleared to make way for agricultural development (Figure 1A). Guanarito virus is vectored by grassland rodents and is typically transmitted to humans through inhalation of virus-infected rodent excrement. The cane mouse (*Zygodontomys brevicauda*) is the principal vector but the virus has also been isolated from cotton rats (*Sigmodon alstoni*). The rodents that vector Guanarito virus have a distribution much broader than that of VHF, and infected rodents have been trapped outside of the VHF endemic region. The incidence of VHF peaks each year from November and January, due to increased agricultural activity. The increased human density and activity in the endemic region during these months may increase the rate of exposure of humans to infected rodents [1].

Different species of arenavirus often overlap in geographic range. The two species of rodents that vector Guanarito virus also vector Pirital virus with cotton rats rather than cane mice acting as the principal vector. Pirital virus is present in the same region of Venezuela but is antigenically distinquishable from Guantarito virus, belongs to a different phylogenetic subgroup of arenaviruses (Figure 2), and has not been associated with human disease [1]. However, cross reactivity between antibodies used to differentiate arenaviruses within the same subgroup (Tacaribe complex) may make it difficult to correctly identify which arenavirus is present in viral hemorrhagic fever cases that occur in regions where two arenavirus species are known to be present [2].

VHF was first reported as an outbreak of 15 cases (9 fatal) in Portuguesa state during 1989. Another outbreak of disease occurred in 1992 resulting in 105 cases and 26 deaths. Sporadic cases occur annually with 18 cases of Venezuelan hemorrhagic fever being reported in 2002 from the states of Portuguesa (15 cases) and Barinas (3 cases) (Figure 1B).

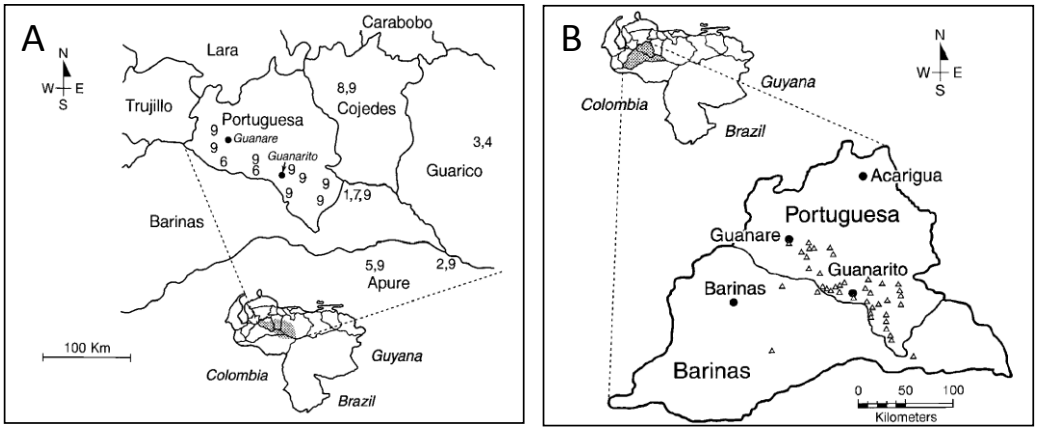


Figure 1. Location of Guanarito virus genotypes (A) and human VHF cases (B). (A) Map of western Venezuela showing the locations where Guanarito virus genotypes (1-9) were isolated[1]. (B) Upper left, map of Venezuela showing the locations (shaded) of the states of Portuguesa and Barinas. Lower right, enlarged map of the same two states, showing the approximate locality of residence (Δ) of all patients with Venezuelan hemorrhagic fever who were reported in 1996 [3].

Viral taxonomy and genome diversity

Guanarito virus is a member of the *Arenaviridae* family, which includes two serocomplexes, the Old World serocomplex and the New World serocomplex. Additionally, the *Arenaviridae* family was been subdivided into five phylogenetic lineages of which Guanarito virus belongs to South American lineage B which includes human pathogens Machupo virus, Junin virus, and Sabia virus (Figure 2).

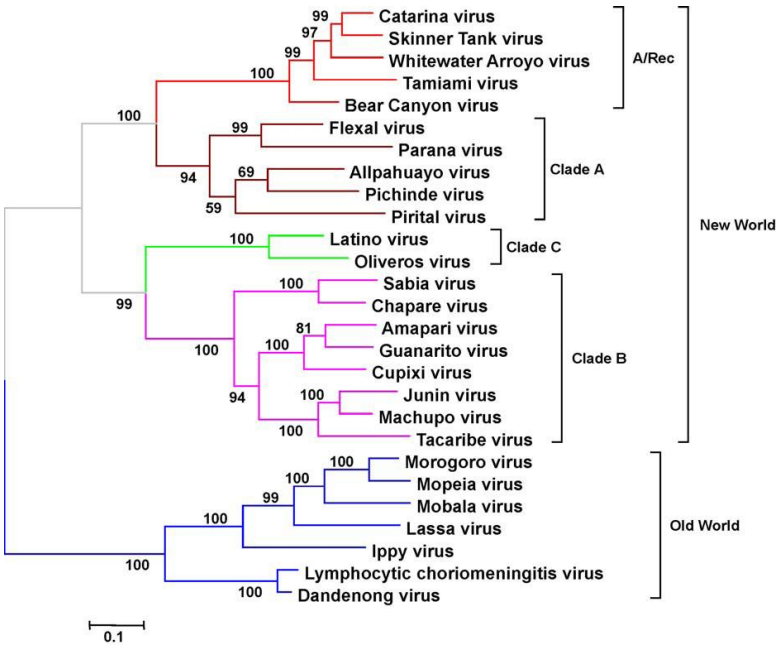


Figure 2. Phylogeny of arenaviruses based on the complete amino acid sequence of the nucleoprotein [4].

Nine Guanarito virus genotypes have been identified via sequence analysis of partial N protein gene from 29 isolates obtained from humans (n=8) and rodents (n=21) between 1990 and 1997. Guanarito virus genotypes differ in nucleotide sequence by 4-10%. One genotype was much more abundant and

widely dispersed as compared to the other eight genotypes (Figure 1A). Only two of the genotypes included isolates from human cases. [1]

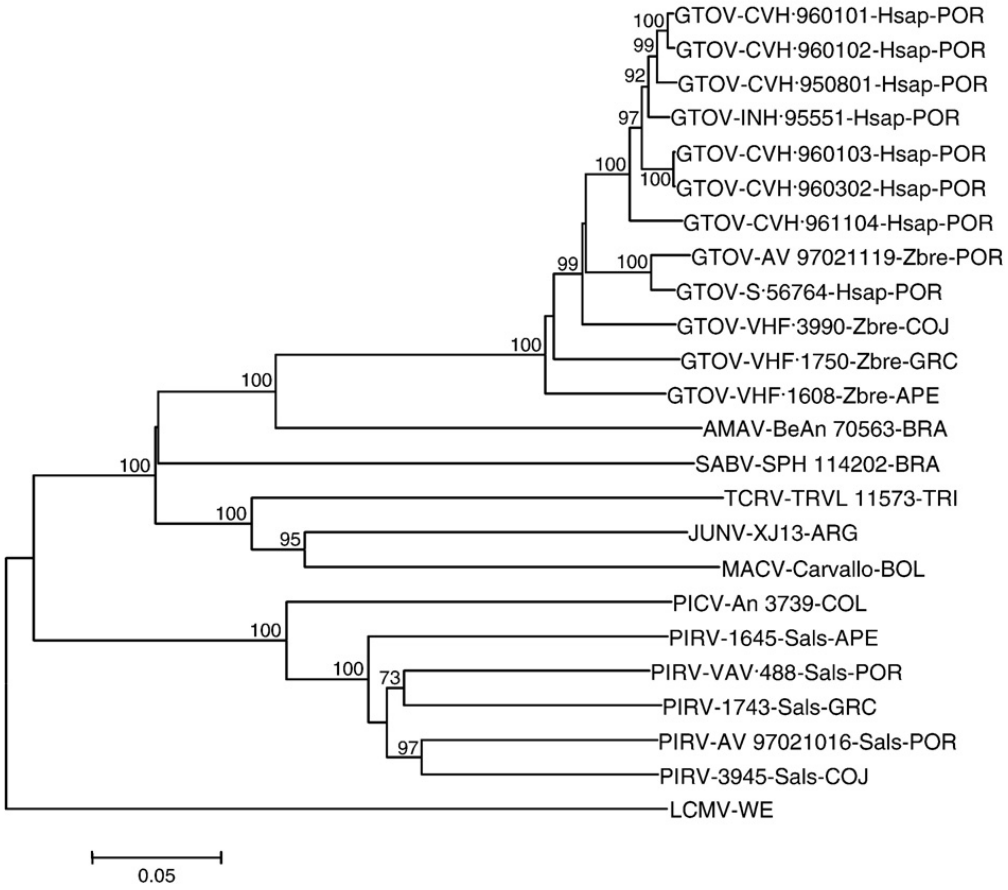


Figure 3. Phylogenetic relationships among 23 South American arenaviruses based on analyses of the complete glycoprotein precursor gene sequences. The branch labels include (in the following order) virus species, strain, and country or state of origin. The branch labels for the GTOV strains and PIRV strains also include the species of the hosts from which these viruses were isolated. The lymphocytic choriomeningitis virus (LCMV) strain WE is an OldWorld arenavirus and was included in the analysis to infer the ancestral node within the group of NewWorld arenaviruses. AMAV, Amapari virus; GTOV, Guanarito virus; JUNV, Junín virus; MACV, Machupo virus; PICV, Pichindé virus; PIRV, Pirital virus; SABV, Sabiá virus; TCRV, Tacaribe virus. Hsap, Homo sapiens; Sals, Sigmodon alstoni; Zbre, Zygodontomys brevicauda. ARG, Argentina; BRA, Brazil; BOL, Bolivia; COL, Colombia; TRI, Trinidad. APE, Apure; COJ, Cojedes; GRC, Guárico; POR, Portuguesa.

Near neighbors of Guanarito virus include Amapari virus, Cupixi virus, Sabia virus, and Tacaribe virus (Figure 2). Amapari virus and Cupixi virus are present in Brazil and are not associated with human illness; these viruses differ from Guanarito virus by 15-31% and 17-31% amino acid difference, respectively. Sabia virus was isolated from a human case of hemorrhagic fever that occurred in Brazil; the geographic distribution and wildlife host of this virus are not known. Sabia virus differs in amino acid sequence from Guanarito virus by 30-48%. Tacaribe virus, a virus isolated from bats in Trinidad, differs from Guanarito virus by 29-47% amino acid sequence difference (Charrel et al., 2002). Whole genome sequence data is available for only one isolate of Amapari virus, Cupixi virus, Sabia virus, and Tacaribe virus.

Gaps in sequence data

The only fully sequenced genome of Guanarito virus was obtained from a human case that occurred in 1990; more recent isolates from humans need to be sequenced. Isolates obtained from infected rodents found inside and outside of the VHF endemic region should also be sequenced. Whole genome sequence data is needed for the near neighbors of Guanarito virus, Amapari virus, Cupixi virus, Sabia virus, and Tacaribe virus, however, it is likely that only a single isolate exists for each of these viruses as there was only one strain identified in searches of scientific literature and in Genbank.

Potential sources of material

Potential sources of strains from human and rodents include Mike Bowen and Stuart Nichol, (CDC); and Bob Tesh and Charles Fulhorst (UTMB).

References

1. Weaver, S.C., et al., Guanarito virus (Arenaviridae) isolates from endemic and outlying localities in Venezuela: sequence comparisons among and within strains isolated from Venezuelan hemorrhagic fever patients and rodents. Virology, 2000. **266**(1): p. 189-95.
2. Cajimat, M.N. and C.F. Fulhorst, Phylogeny of the Venezuelan arenaviruses. Virus Res, 2004. **102**(2): p. 199-206.
3. de Manzione, N., et al., Venezuelan hemorrhagic fever: clinical and epidemiological studies of 165 cases. Clin Infect Dis, 1998. **26**(2): p. 308-13.
4. Emonet, S.F., et al., Arenavirus genetic diversity and its biological implications. Infect Genet Evol, 2009. **9**(4): p. 417-29.

Hemorrhagic fever-associated arenaviruses for which the animal vector is unknown: Sabia virus, Chapare virus, and Lujo virus

Summary

Three arenavirus species (Sabia virus, Chapare virus, and Lujo virus) have been isolated from human hemorrhagic fever cases but have not yet been detected in an animal vector, therefore the geographical range and ecology of these viruses remains undetermined. In each case a single complete genome sequence is available because either only one clinical isolate was available (Sabia virus and Chapare virus), or multiple samples obtained from different cases from a single outbreak were determined to be genetically identical [1-3]. Provided follow up studies are done on potential rodent vectors and/or additional human cases occur, future sources of isolates may include Stuart Nichol (CDC) and the World Reference Center of Emerging Viruses and Arboviruses (UTMB).

Ecology and epidemiology

Sabia virus emerged in 1990 when it caused a fatal case of hemorrhagic fever that occurred in Sao Paulo, Brazil (Figure 1A). The disease was initially believed to be caused by yellow fever virus, however the virus isolated proved to be an antigenically and genetically distinct arenavirus [4]. Subsequently the virus caused nonfatal influenza-like illness in two laboratory technicians who were exposed while working

with the virus [5]. No rodent vector has been identified for Sabia virus. Amapari and Cupixi arenaviruses are also present in Brazil but are not associated with human illness.

Chapare virus was isolated from a fatal case of hemorrhagic fever that was part of a small outbreak that occurred between December 2003 and January 2004 near Cochabamba, Bolivia [2]. According to the EmergingDisease.org web site (http://www.satolah.com/birdflu/New%20Virus%20(Arenavirus)%20Causes%20South%20American%20Fever.pdf)http://www.satolah.com/birdflu/New Virus (Arenavirus) Causes South American Fever.pdf), several rodents were trapped and tested for this virus but were found to be negative. The site also quoted Stuart Nichol as mentioning that political issues may make the region difficult to access. Stuart Nichol (CDC) or Tom Ksiazek (UTMB) may have information regarding follow-up studies on the ecology of this virus.

Lujo virus was identified in South Africa in 2008 as the cause of a small outbreak of hemorrhagic fever that originated in Zambia (Figure 1B). This outbreak involved secondary and tertiary spread and had a mortality rate of 80% (4/5 cases) [1]. The index case was transferred from Zambia to South Africa via air transport and contact between the index case and medical personnel initiated nosocomial transmission.

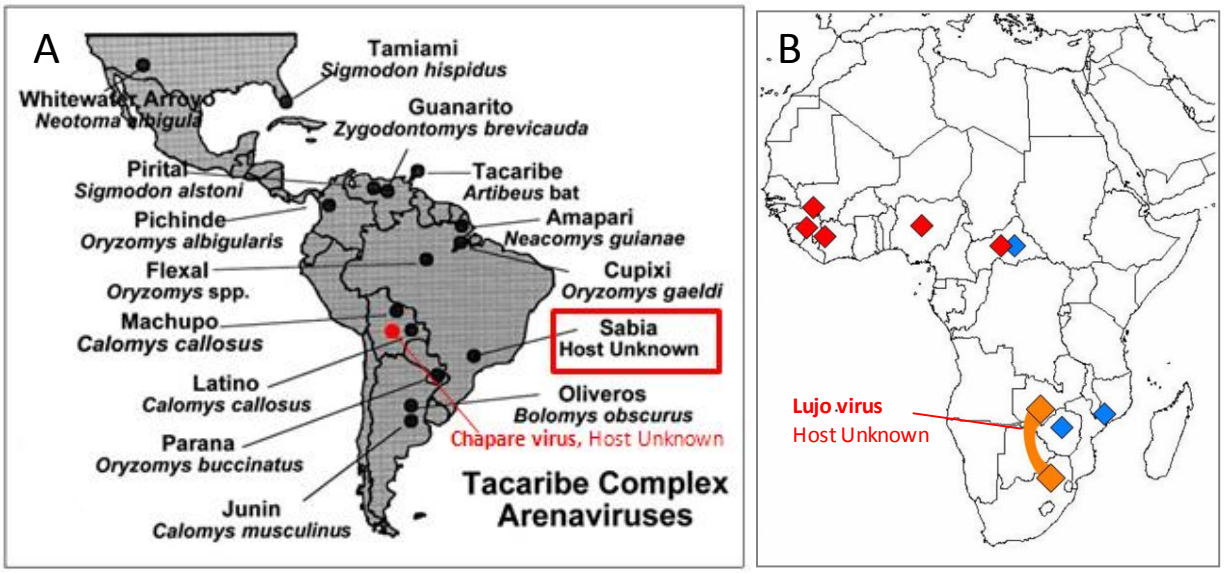


Figure 1. Comparison of the geographic range of American arenaviruses and African arenaviruses. (A) Location of American arenaviruses and their natural hosts (figure adapted from Fields Virology, 2007). (B) Geographic distribution of African arenaviruses. Mobala virus, Mopeia virus, and Ippy virus (blue) have not been implicated in human disease; Lassa virus (red) can cause hemorrhagic fever. The origin of the newly discovered arenavirus, Lujo virus, index and secondary and tertiary cases linked in the 2008 outbreak are indicated in gold [1].

Viral taxonomy and genome diversity

The *Arenaviridae* family includes two serocomplexes, the Old World serocomplex and the New World serocomplex, which also known as the Tacaribe complex. Sabia virus and Chapare virus are members of the Tacaribe complex, Clade B, and are most closely related to each other with 26% and 30% nucleotide sequence difference for the S segment and the L segment respectively (Figure 2) [2]. Sabia virus differs in amino acid sequence from near neighbors Chapare virus, Machupo virus, Cupixi virus, and Junin virus by 15-26%, 27- 50%, 22- 41%, and 29-45% depending on the gene sequence analyzed. The amino acid

differences for Chapare virus and near neighbors were similar to those of Sabia virus (Supporting Table 1, [2]).

Phylogenetic analysis of the Lujo virus genome placed it as a branch off the base of the Old World serocomplex, distinct from the LCMV lineage that includes Lassa virus (Figure 2) [1]. Near neighbors of Lujo virus include Ippy virus (36% nucleotide difference), and Mopeia virus and Mobala virus (37% nucleotide difference each) (Supporting Table 1, [1]).

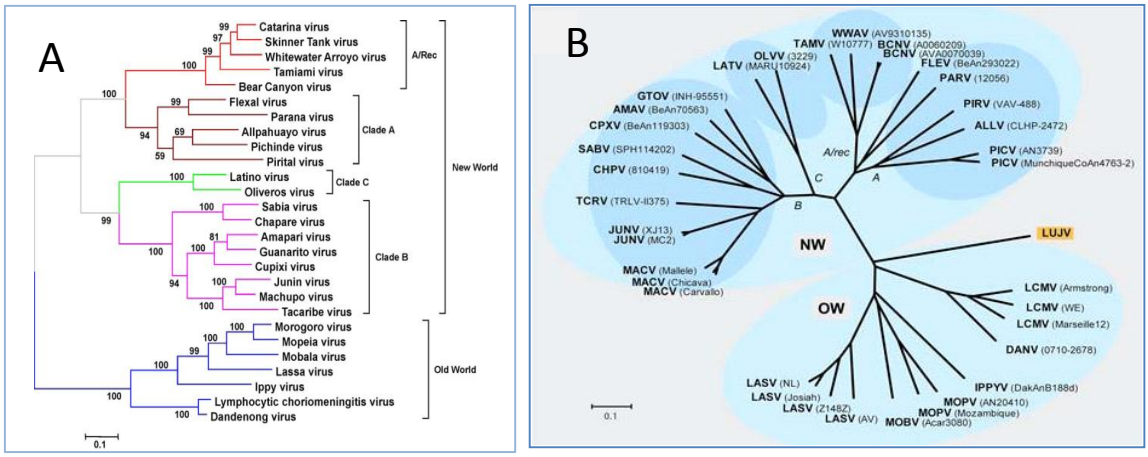


Figure 2. Phylogeny of New World and Old World arenaviruses. (A) Phylogenetic tree of arenaviruses based on the complete amino acid sequence of the nucleoprotein prior to the discovery of Lujo virus [4]. (B) Analysis of complete sequences of S segment shows that Lujo virus, branches off the root of the Old World arenaviruses, and may be a novel genetic lineage, separate from the lymphocytic choriomeningitis virus (LCMV) lineage which includes Lassa virus (LASV), Mopeia virus (MOPV), and Mobala virus (MOBV) [1].

Gaps in sequence data

Genetic and ecological knowledge for each of these viruses is limited to that obtained from a single human outbreak. Until these viruses are detected in an animal (most likely rodent) vector the potential for future outbreaks cannot be predicted. Similarly, the small number of clinical samples available for genetic analysis limits the robustness of assays developed for detection of these viruses. Whole genome sequence data is needed for the near neighbors of Sabia virus, Chapare virus, and Lujo virus including Cupixi virus and Ippy virus. However, it is likely that only a single isolate exists for each of these viruses as there was only one strain identified in searches of scientific literature and in Genbank.

Potential sources of material

Potential sources of strains from human and rodents include Stuart Nichol, (CDC); and Bob Tesh, Tom Ksiazek and Charles Fulhorst (UTMB).

References

1. Briese, T., et al., Genetic detection and characterization of Lujo virus, a new hemorrhagic fever-associated arenavirus from southern Africa. PLoS Pathog, 2009. **5**(5): p. e1000455.
2. Delgado, S., et al., Chapare virus, a newly discovered arenavirus isolated from a fatal hemorrhagic fever case in Bolivia. PLoS Pathog, 2008. **4**(4): p. e1000047.

3. Gonzalez, J.P., et al., Genetic characterization and phylogeny of Sabia virus, an emergent pathogen in Brazil. Virology, 1996. **221**(2): p. 318-24.
4. Lisieux, T., et al., New arenavirus isolated in Brazil. Lancet, 1994. **343**(8894): p. 391-2.
5. Charrel, R.N. and X. de Lamballerie, Arenaviruses other than Lassa virus. Antiviral Res, 2003. **57**(1-2): p. 89-100.

Melioidosis: *Burkholderia pseudomallei*
Glanders: *Burkholderia mallei*

Summary

Only 11 complete genome sequences of *Burkholderia mallei (B. mallei)* exist, which is half the number of completely sequenced genomes than for any other bacterial biothreat agent. Additional *B. mallei* genomes should be sequenced, especially from poorly-represented geographic locations such as South America and Africa. Of the completely sequenced genomes of *Burkholderia pseudomallei (B. pseudomallei)* where a species of isolation was identified, all were isolated from humans. It is desirable to sequence additional genomes isolated from other hosts. Australia is thought to be the source of origin of *B. pseudomallei* as isolates from this continent are most genetically diverse.[1] Since Thailand may potentially have isolates that are more virulent to humans[2] additional human isolates from Thailand should also be sequenced. There are four completely sequenced genomes (one finished and three shotgun sequences) from *Burkholderia thailandensis (B. thailandensis),* the closest genetic near neighbor of *B. mallei* and *B. pseudomallei.* Additional genomes need to be sequenced. A great deal of sequence information is accumulating for the next closest genetic neighbor (*B. cepacia*) so no additional sequences of this near neighbor are needed. A tremendous amount of genetic diversity exists within *B. pseudomallei* including recently identified genomic island differences[3], additional whole genome comparisons between sequenced genomes of *B. pseudomallei* are warranted as are additional studies to determine the parent progeny relationship of *B. thailandensis, B. mallei* and *B. pseudomallei.*

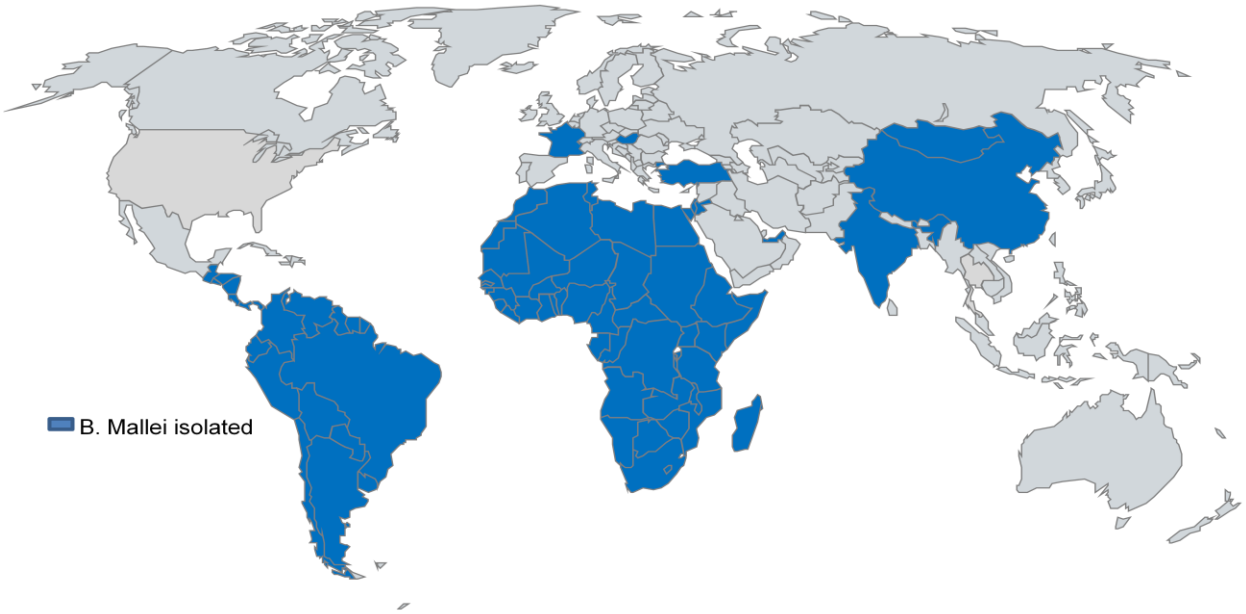


Figure. 1. Global map indicating countries where *B. mallei* is found.

Ecology and epidemiology

Melioidosis is caused by the bacteria *Burkholderia pseudomallei* while *Burkholderia mallei* causes the disease glanders. While *B. mallei and B. pseudomallei* are quite close phylogenetically their ecology and epidemiology are quite divergent. *B. mallei* a potential human pathogen is typically a pathogen of horses, mules and donkeys. Humans are typically infected through contact with infected animals or animal products. *B. mallei* can be found in the animal populations in Asia, the Middle East, Central and South America and Africa (Fig. 1). Since glanders was eradicated from North America and Western Europe around 1950 there have been no cases of glanders in the US in sixty years.

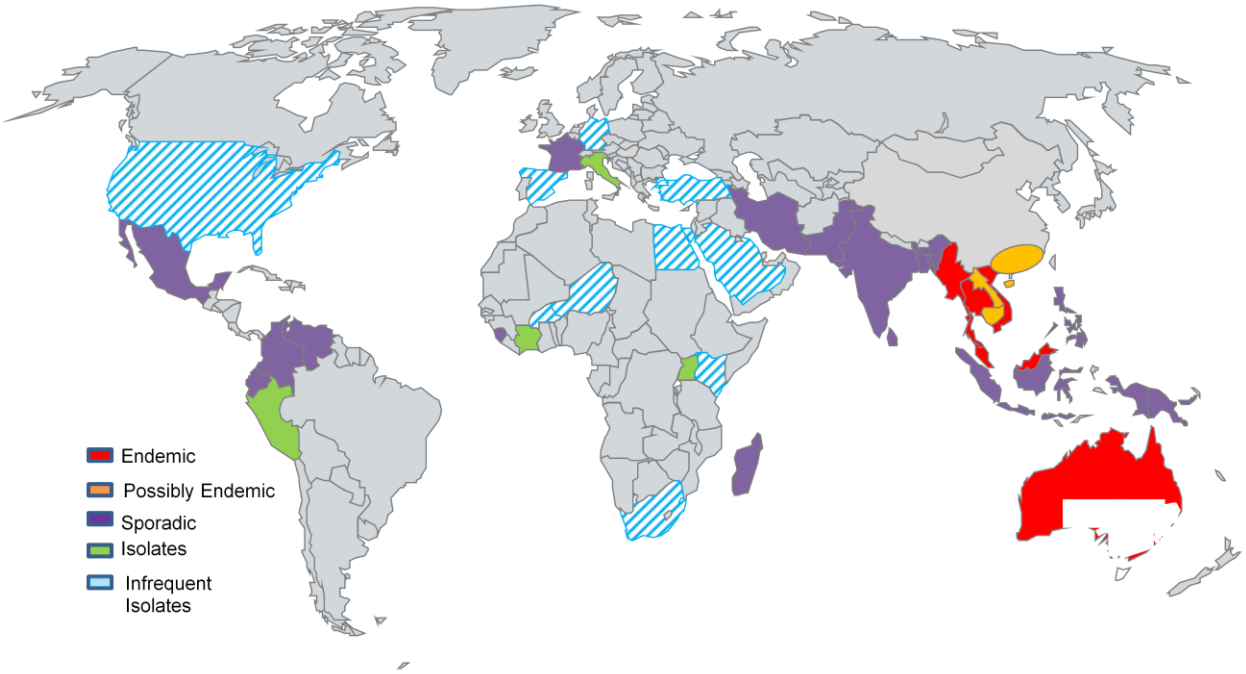


Fig. 2. Global map indicating geolocation of B. *pseudomallei* adapted from Ref. 4.

B. pseudomallei is a human pathogen endemic to Southeast Asia and Northern Australia although it has been spread elsewhere (Fig. 2). Additional genomes from South America and Africa should be sequenced to help define the global relatedness and dispersion of *B. pseudomallei i*solated from outside of Australia and Southeast Asia. *B. pseudomallei* lives on dead or decaying matter and can be cultured from soil or ground water and is typically contracted from extended contact with contaminated soil and water.

Both glanders and melioidiosis can have acute or chronic forms of human infection. In an acute infection, symptoms include fever, lethargy, abscess formation, pneumonia and sepsis. Septicemia caused by *B. pseudomallei* can have a mortality of forty percent even with antibiotic therapy. Since cases of human glanders in the US are exceedingly rare not as much information exists regarding survival rates for individuals with *B. mallei* infections.[5] While glanders is a disease primarily restricted to horses, mules, donkeys with occasional human infection, domestic animals such as sheep, pigs, goats, cattle, horses, camels can be infected with *B. pseudomallei* as well as wild animals such as primates, dolphins, marsupials, birds, reptiles and fish. However a recent ProMed search identifies only one report of

melioidiosis (Australia) and two of glanders (Russia, Iran) reported in the last three years and none of these in humans.

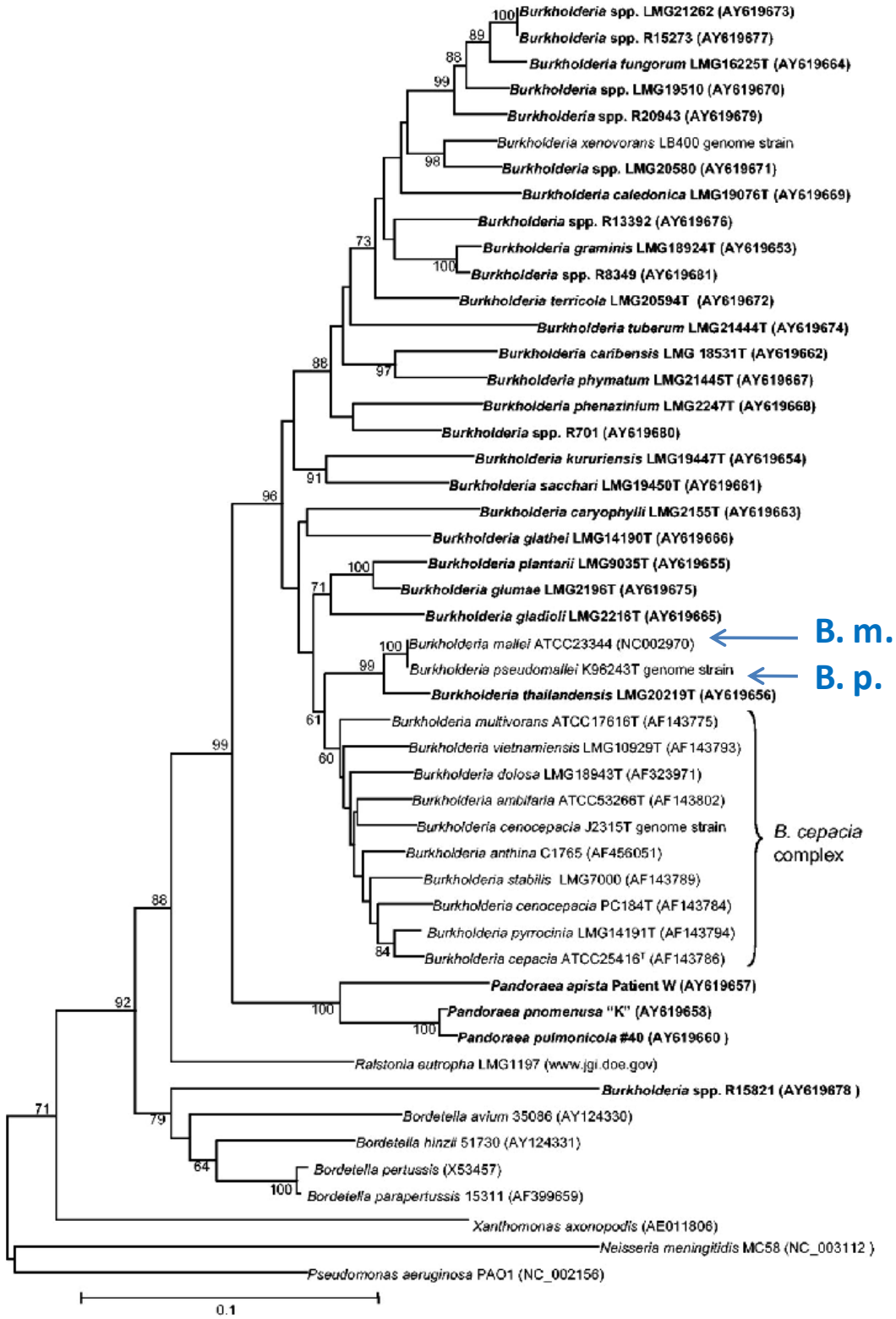


Fig. 3. Phylogenetic tree of *Burkholderia* genus using *recA* sequences (from Ref. 6). Tree is rooted using *P. aeruginosa recA* gene.

Bacterial taxonomy and genome diversity

Unlike *B. anthracis, Y. pestis* and *F. tularensis* the genome of *B. mallei* and *B. pseudomallei* consists of two chromosomes and no plasmids. The *B. mallei* genome is the smaller of the two genomes with a total of 5.7 Mb with one chromosome of 3.5 and the other of 2.3 Mb. The genome of *B. pseudomallei* is a total of 7.2 Mb with two chromosomes of 4.07 and 3.17 Mb in size. While the large chromosome primarily harbors genes essential for growth, the small chromosome contains more diverse genes that are primarily involved in survival or exploiting variable environmental conditions. *Burkholderia* is a genus with complex taxonomy that contains more than 40 species. *Burkholderia* species are widely distributed in the environment some as free living symbionts and commensals and some as pathogens of both plants and animals.

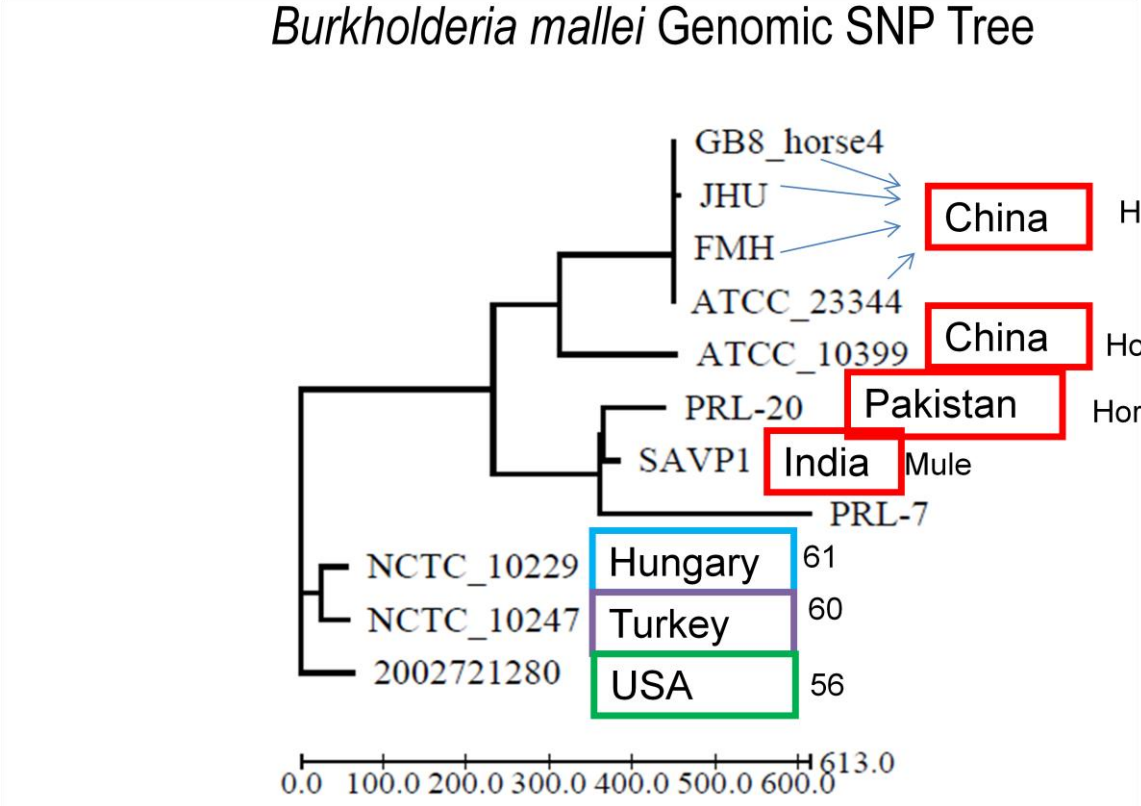


Fig. 4. Phylogenetic SNP alignment of completely sequenced *B. mallei* genomes.
Americas, Europe, Middle East, Asia.

Figure 3 shows the relationship between *B. mallei* and *pseudomallei* and other members of the genus *Burkholderia*. This tree is based upon a single gene, due to a lack of sufficient whole-genome data. The positions of *B. mallei and B. pseudomallei* on the tree are indicated by blue arrows. *B. mallei* (the only obligate pathogen) is considered to be a clonal branch of *B. pseudomallei* [7]with all isolates of *B. mallei* falling in a genetically restricted group within the genetic diversity of *B. pseudomallei.* The closest genetic near neighbor of *B mallei* and *B. pseudomallei* is *B. thailandensis* which is a nonpathogenic bacteria. *B. cepacia* which is a pathogen for people suffering from cystic fibrosis is a complex of nine species of *Burkholderia* (Fig. 3).

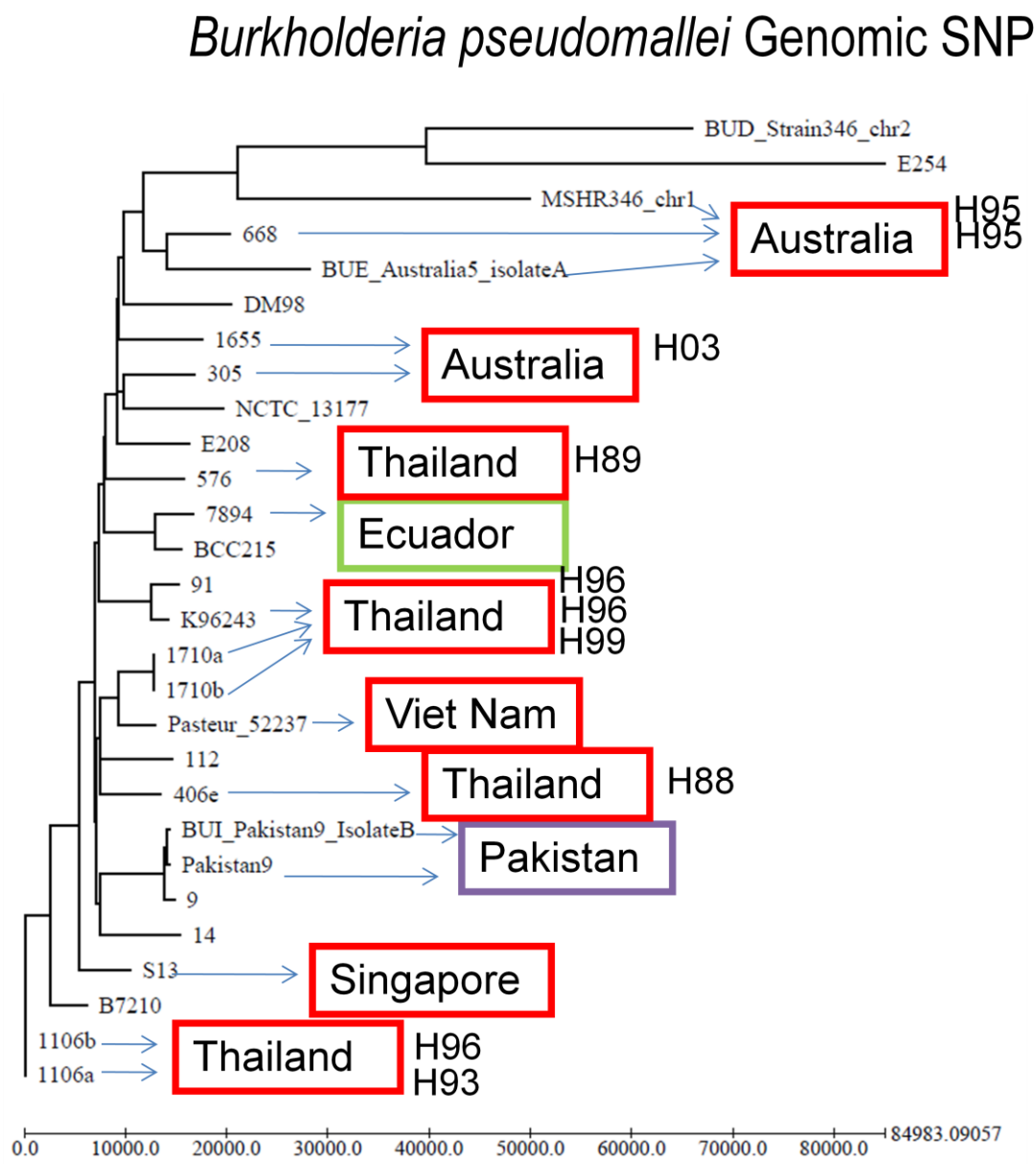


Fig. 5. Phylogenetic SNP alignment of completely sequenced *B. pseudomallei* genomes.
Americas, Europe, Middle East, Asia.

While sequences of genomes of *B. cepacia* complex species are rapidly accumulating, there are only four complete genomes of *B. thailandensis.* Additional genomes of *B. thailandensis* need to be sequenced as it is considered by some to be the parent of *B. pseudomallei* from which in turn *B. mallei* is thought to be derived.[7]

Figure 4 illustrates a phylogenetic SNP tree of the completely sequenced genomes of *B. mallei* (Gardner et al. unpublished LLNL bioinformatics). The country of isolation is shown as a colored box. For those isolates for which a date of isolation for the genome was found, the date is listed as a two digit number to the right of the country of origin box. In those cases where a species of origin was found, it is indicated by an H if human or the name of the species is listed between the two digit year and the

country of isolation box. There are only 11 completely sequenced genomes of *B. mallei* which unlike *B. pseudomallei* is an obligate mammalian pathogen. Of the 11 the four at the top of the tree from China are clones derived from a single infection. More genomes of *B. mallei* need to be sequenced.
Fig. 5 shows a phylogenetic SNP tree (Gardner et al. LLNL bioinformatics) of the completely sequenced genomes of *B. pseudomallei*. The country of isolation is shown as a colored box. For those isolates for which a date of isolation for the genome was found it is listed as a two digit number to the right of the country of origin box. Information for a species of isolation could be found for 10 of 27 sequenced genomes. However, all of those with an identified species are human cases. As noted earlier, *B. pseudomallei* can infect a diverse range of domestic and wild animals and genomes isolated from additional animal species should be sequenced. While *B. pseudomallei* is endemic in Australia and Thailand, additional isolates from Africa and South America should be sequenced to determine the geographic relationships between isolates from other regions. Additional environmental isolates from Thailand are needed to help define genetic diversity for *B. pseudomallei* in this region and because the most pathogenic human pathogenic isolates are found in Thailand.[2] Additional isolates of *B. pseudomallei* from Australia are also needed as this continent has isolates of the greatest genetic diversity and Australia is believed to be the site of origin of *B. pseudomallei*.[1]

New genetic diversity in *B. pseudomallei* isolates was identified from a comparison of completely sequenced genomes. For example, two mutually exclusive genomic island (GI) gene cassettes, termed "BTFC and YLF", have been described that are dissimilar in their geographical distribution. The BTFC cassette is found only in *B. pseudomallei* genomes isolated from Australia and in *B. thailandensis* which only been isolated in Thailand. However the YLF cassette has been found only in *B. pseudomallei* genomes isolated from Thailand and there is some correlation between this cassette and increased pathogenicity.[8]

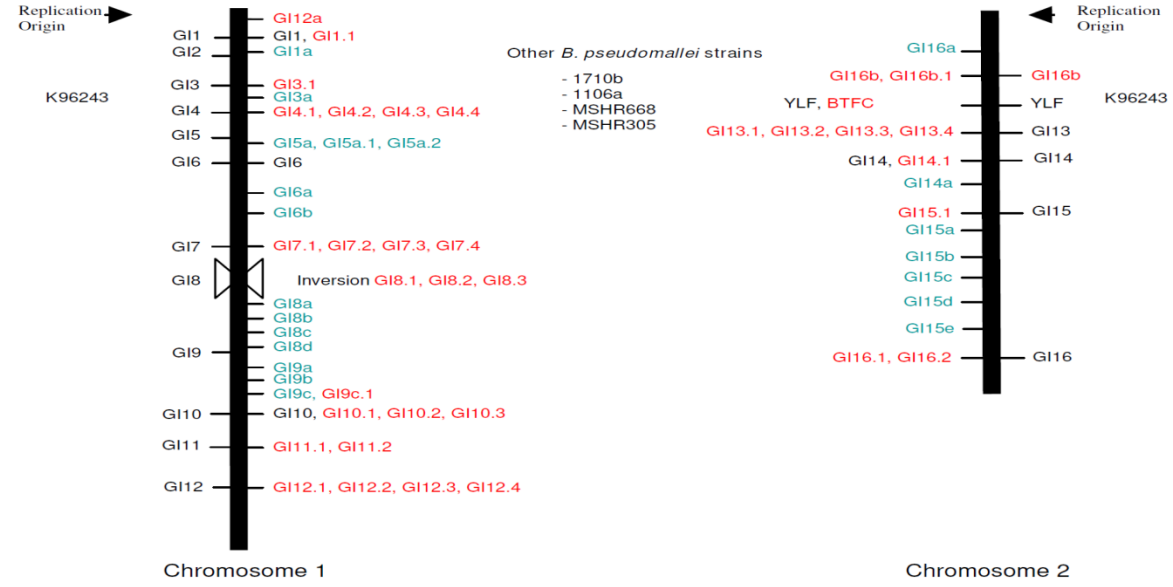


Fig. 6. Genomic locations of 71 GIs on chromosomes 1 and 2 in *B. pseudomallei*. GIs from strains 1710b, 1106a, MSHR668, MSHR305 (inside dark chrom. lines) compared to 16 GIs from K96243 (outside lines) from Ref. 3.

The genomic diversity in *B. pseudomallei* was shown to be even more complex by the recent comparison of five completely sequenced genomes (K96243, 1710b, 1106a, MSHR668 [668 in Fig. 5], and MSHR305

[305 in Fig.]). Seventy-one distinct GIs were identified in the sequences from these five genomes (Fig. 6).

Since this diversity was discovered in a comparison of only five completely sequenced genomes out of twenty seven, a great deal of additional bioinformatics comparisons need to be done between completely sequenced genomes to determine the extent of this newly discovered genetic diversity. These genomic islands range in size from 4 to 100 kb and contain sets of genes which can dramatically change the pathogenic phenotype of *B. pseudomallei*. In addition to the sequencing of new genetically and geographically diverse isolates of *B. pseudomallei*, work obviously remains to be done in the genomic comparison and analysis of the existing set of completely sequenced genomes of *B. pseudomallei*. A further comparison of *B. pseudomallei* to *B. mallei* genomes also needs to be done in light of this new GI information as *B. mallei* is an obligate mammalian pathogen.

Gaps in sequence data

Virulence and Resistance

Since *B. pseudomallei* is believed to be a soil microorganism its environment requires the presence of multiple efflux pumps that also provide multiple mechanisms for antibiotic resistance. To further complicate the picture the recently discovered wealth of diversity in pathogenicity islands in *B. pseudomallei* give it the potential for additional unknown genetic diversity with respect to both virulence factors and antibiotic resistance.[3] *B. pseudomallei* movement from the environment as well as attachment and invasion of epithelial and macrophage cells and intracellular survival and spread has been recently reviewed.[9] A diverse assortment of virulence factors are known to allow *B. pseudomallei* to become an effective opportunistic pathogen, as well as to avoid or subvert the host immune response. Genomic and molecular studies have substantially increased the current understanding of the infection process of melioidosis yet, much of these processes still remain to be elucidated. The limited amount of sequence information and isolates for *B. mallei* makes the knowledge of virulence mechanisms for this pathogen even less well understood than for *B. pseudomallei.* The other attribute of *B. mallei* and *B. pseudomallei*, namely their resistance to antibiotics, has been determined for 65 isolates.[10] On the other hand, the virulence and antibiotic resistance of *B. mallei* is not well understood because there are so few human infections and so few genomes sequenced.

Geolocation

Additional genomes of *B. mallei* from throughout the globe need to be sequenced as there is only a handful of completely sequenced genomes for this zoonotic pathogen. Additional environmental isolates from Australia, the putative site of origin of *B. pseudomallei,* as well as clinical isolates from Thailand (where the disease has greatest pathogenicity) need to be sequenced. Additional isolates from Africa, South America and isolates similar to *B. oklahomensis* from USA need to be sequenced in order to better define the geographic distribution of *B. pseudomallei*.

Additional Species

Given that *B. pseudomallei* is known to infect a wide variety of domestic and wild animals and that of the complete genomes for which as species of isolation has been found all are from humans, more isolates from other hosts should be sequenced.

Phylogeny

It is believed that *B. mallei* is derived from *B. pseudomallei* and that possibly *B. pseudomallei* derived from *B. thailandensis* but more sequence information is needed to determine these genetic relationships. Recent identification of large genomic islands of sequence difference between five completely sequenced genomes of *B. pseudomallei* needs to be expanded to include all the completely sequenced genomes of *B. pseudomallei* and extended to *B. mallei.* The impact of gene differences found in these GIs on the phenotype (for example human pathogenicity or antibiotic susceptibility) of different strains of *B. pseudomallei* may be considerable.

Genetic Near Neighbor

Additional isolates of *B. thailandensis* should be sequenced as there are only four sequenced genomes of this closest genetic near neighbor of *B. mallei* and *B. pseudomallei.* While a wealth of sequence information is accumulating for *B. cepacia* additional genetic comparisons should be done between this genetic near neighbor and *B. mallei* and *pseudomallei* to determine genetic relationships and human virulence factors.

Potential sources of material include:

NBFAC (Jim Burans)
NAU (Paul Keim)
USDA NVSL (Matt Erdman, Linda Schlater)
BYU (Rich Robison)
LANL (Cheryl Kuske)

REFERENCES

1. VNTR analysis of selected outbreaks of *Burkholderia pseudomallei* in Australia. Pearson, T., et al., Infection, Genetics and Evolution. 2007, 7:p416.

2. Nonrandom Distribution of *Burkholderia pseudomallei* Clones in Relation to Geographical Location and Virulence. Journal of Clinical Microbiology. Vesaratchavest, M., 2006, 44:p2553.

3. Genomic islands from five strains of *Burkholderia pseudomallei.* Tuyanyok, A., BMC Genomics. 2008, 9:p566.

4. Environmental Factors That Affect the Survival and Persistence of *Burkholderia pseudomallei.* Inglis T.J. Applied and Environmental Microbiology. 2006, 72:p6865.

5. Strategies toward vaccines against Burkholderia mallei and Burkholderia pseudomallei. Bondi, S.K., et al. Expert Review of Vaccines. 2008, 7:p1357.

6. Development of a *recA* Gene-Based Identification Approach for the Entire *Burkholderia* Genus. Payne, G.W. et al., Applied and Environmental Microbiology. 2005, 71:p3917.

7. Tandem repeat regions within the *Burkholderia pseudomallei* genome and their application for high resolution genotyping. U'Ren, J.M., et al. BMC Microbiology. 2007, 7:p23.

8. A horizontal gene transfer event defines two distinct groups within Burkholderia pseudomallei that have dissimilar geographic distributions. Tuanyok, A. et al., Journal of Bacteriology. 2007, 189:p9044.

9. The molecular and cellular basis of pathogenesis in melioidosis: how does Burkholderia pseudomallei cause disease? Lazar Adler, N.R., et al., FEMS Microbiology Reviews. 2009 Epub.

10. Antibiotic susceptibility of 65 isolates of *Burkholderia pseudomallei* and *Burkholderia mallei* to 35 antimicrobial agents. Thibault, F.M., et. Al., Journal of Antimicrobial Chemotherapy. 2004, 54:p1134.

Tularemia: *Francisella tularensis*

Summary

In a 2004 publication Johansson[1] noted that the limited number of *Francisella tularensis* (*F.t*.) isolates from Japan and a dearth of isolates representing subspecies *mediasiatica* prevented an exhaustive analysis of *Francisella* diversity. This gap in sequence information still exists in 2009. Additional samples of *F.t. mediasiatica* should be sequenced since there is only one completely sequenced genome. Additional isolates of *F. t. holarctica* from Japan similar to FSC 022 genome (Fig. 3) should be sequenced as it is quite genetically divergent from other *holarctica* strains. Isolates from California (subclade B.Br.002/003) of subspecies *holarctica* should be sequenced as they map genetically between FSC 022 and all other isolates of the *holarctica* subspecies.[2] The genetic relationship of *novicida* as a subspecies of *F. t.* or species within the *Francisella* genus may be clarified by sequencing additional divergent isolates of *F.t. novicida* from Australia which have been shown to cause human infection. The relationship of the *F. t.* species to other genetic near neighbor species within the phylogenetic tree (*F. philomiragia*) depends on sequencing new soil isolates[3] (& Fig. 2), clinical isolates that may be new *Francisella* species[4], and additional genomes of *Wolbachia* (tick endosymbionts) and *F. piscicida* (Fig. 2).

Ecology and epidemiology

Tularemia is caused by subspecies of *F. tularensis* (*F.t*.) although the severity of the disease varies a great deal depending on the subspecies. The three recognized subspecies of *F. t.* are *tularensis*, *holarctica*, *and mediasiatica. Francisella novicida* has been classified both as a separate species or subspecies of *F.t*. Of the three subspecies *tularensis* (also known as type A) causes the most severe disease in humans, *holarctica* (aka type B) causes much milder disease, the pathogenicity of *mediasiatica* is not clear as there are few isolates. *F. novicida* is not generally pathogenic to humans. The global distribution of each subspecies (or species) is shown in Fig. 1. With the exception of an isolate from Australia, *F. tularensis* is confined to the northern hemisphere. In addition to *F. novicida, F. tularensis* is also closely related to another *Francisella* species *F. philomiragia*. The ecology of *Francisella* species and subspecies is still poorly understood with many knowledge gaps regarding reservoirs and vectors.

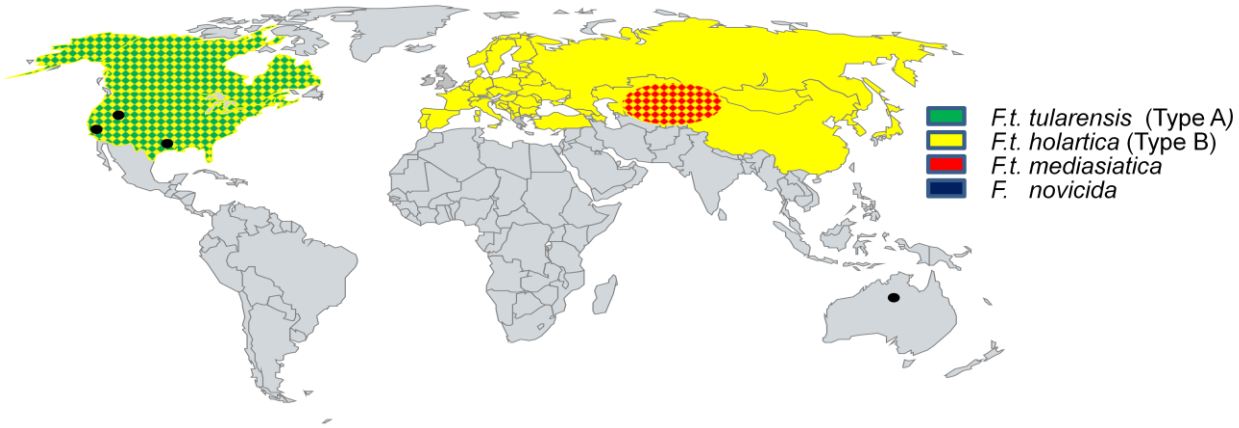


Fig. 1. Global map of prevalence of *F. t.* subspecies *tularensis, holarctica, mediasiatica* and *F. novicida* adapted from Ref. 5.

Tularemia can be transmitted to man by insect bite or by handling and dissection of infected rabbits and rodents.[5] However *Francisella* is found in more than 150 species of animals, including birds and insects, worldwide. Ingesting infected water, soil, or food can also cause *F. tularensis* infection. Depending on the site of infection, tularemia has six characteristic human clinical syndromes: ulceroglandular (the most common type representing 75% of all forms), glandular, oropharyngeal, pneumonic, oculoglandular, and typhoidal. A ProMED search of reports of tularemia cases for the last three years indicates cases from Germany, Sweden, Georgia, Armenia, Norway, Spain, Russia, USA states of AK, ID, CO, SD, UT, TX, NJ, NM,NY, NV, MA, WY.

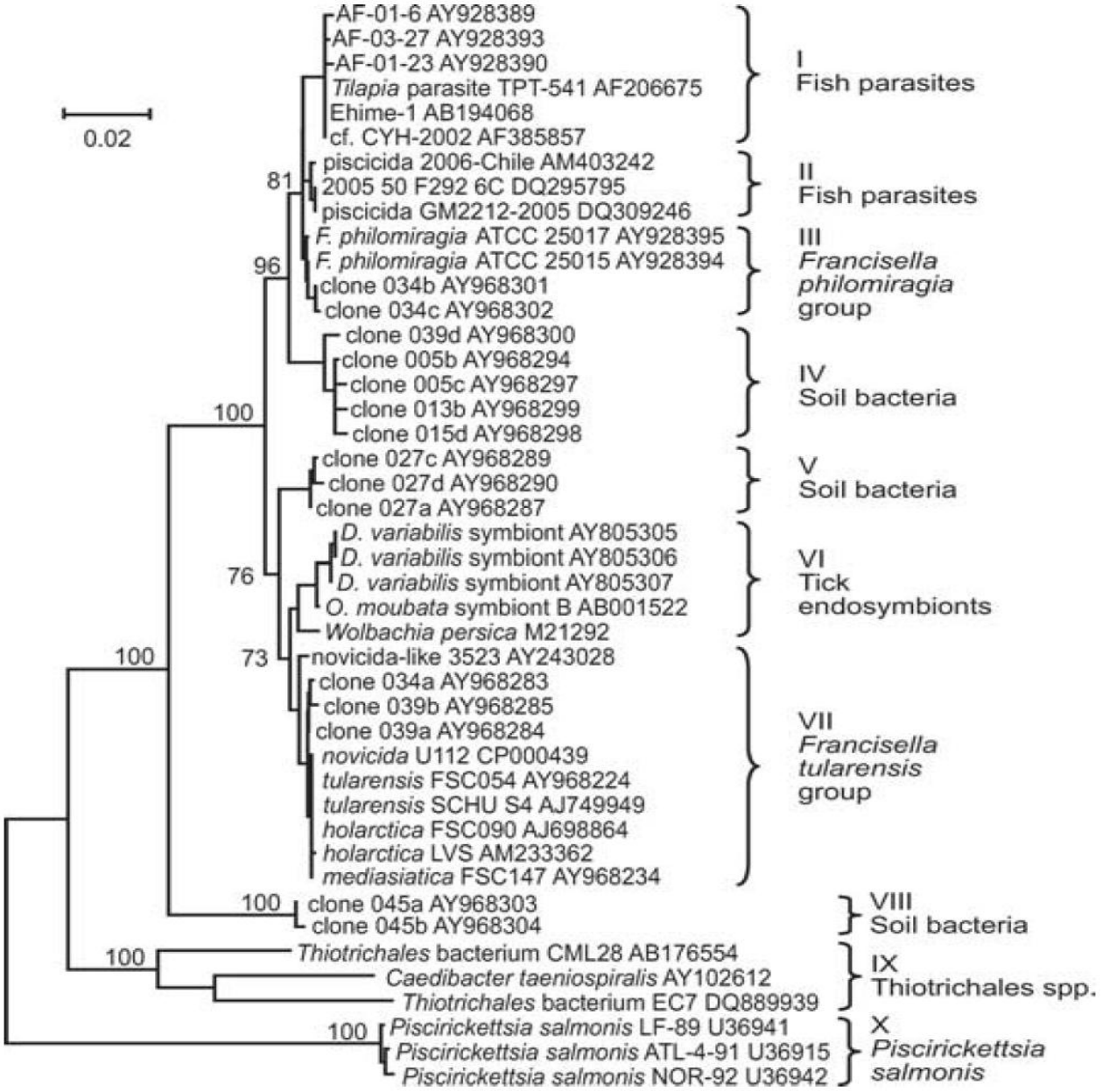


Fig. 2. Phylogeny based on 16S RNA sequence of *Francisella* and representative genetic relatives (from Ref. 5). Isolates labeled soil bacteria are samples collected as a follow-up to environmental aerosol sampling (Ref. 3).

Bacterial taxonomy and genome diversity

The genome of *F.tularensis* is smaller than that of most other biothreat bacterial genomes at 1.9 Mbp and *tularensis*, *holarctica* and *mediasiatica* subspecies have not been found to contain plasmids. In contrast *F. novicida* species have been shown to contain plasmid pFNL10 (4kb).[6] Recently *F. philomiragia* has also been shown to contain two plasmids pF242 and pF243 where pF243 is closely related to pFNL10.[6]

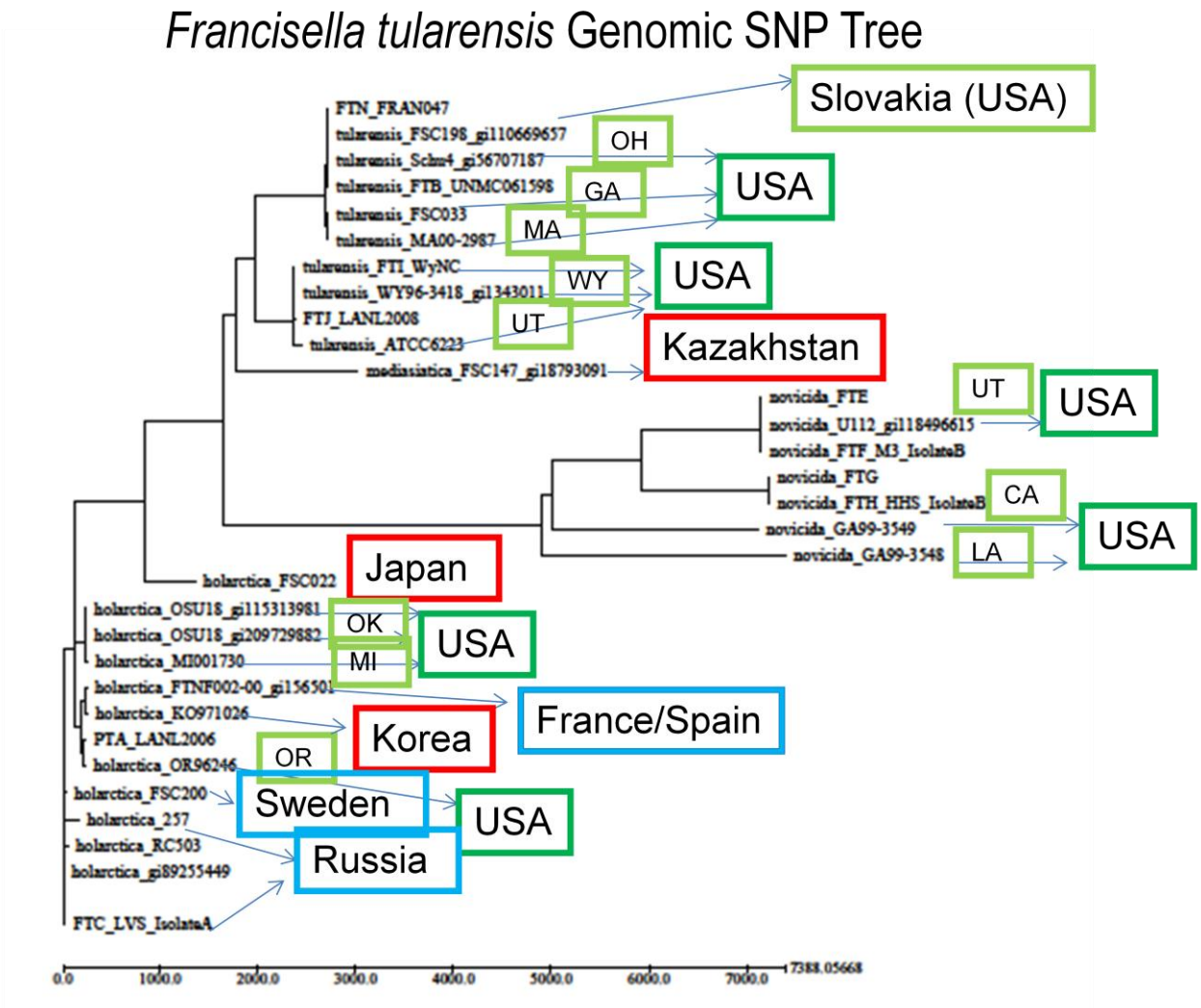


Fig. 3. Phylogenetic SNP alignment of completely sequenced FT genomes (Gardner unpublished, LLNL). Americas, Europe, Africa, Asia.

Figure 2 shows the relationship between *Francisella tularensis* and other related bacteria. Although *Francisella tularensis* and *philomiragia* species are well known due to their close genetic relationship, a large number of close genetic relatives exist in a wide variety of environments including soil (IV, V, VIII), fish (I, II) and tick endosymbionts (VI). The genetic diversity and ecological reservoirs of *F.t.* genetic near neighbors is almost completely unknown. This is due to the limited number of attempts that have been made to do comprehensive environmental surveys of areas where *Francisella* is endemic and until recently[7] no selective culture methods existed for environmental *Francisella* isolation. Additional isolates of from groups II, VI, and VII in Fig. 2 should be sequenced if possible as well as recent clinical isolates that may be entirely new species of human pathogenic *Francisella*.[4]

Figure 3 contains a phylogenetic SNP tree (Gardner et Al. LLNL bioinformatics) of the completely sequenced genomes of *Francisella novicida* as well as *F.t.* subspecies *tularensis*, *mediasiatica* and *holarctica*. The country of isolation is shown as a colored box. For those isolates from the USA, the state from which it was isolated is shown as a light green box along the arrow connecting the genome to the USA box. Of particular interest is an isolate from Slovakia that was found to be genetically very similar to

the Shu 4 strain isolated from Oklahoma. This isolate is listed on the tree with site of origin from both USA and Slovakia. The smaller amount of genetic diversity in subspecies *holarctica* when compared to *tularensis* or *novicida* is indicated by shorter horizontal line lengths between genomes of *holarctica* when compared to *tularensis* or *novicida*. The exception to the monophyletic nature of *holarctica* is FSC 022 Japanese isolate of *holarctica* which is divergent both from other *holarctica* as well as from *novicida*, *tularensis* and *mediasiatica*. Additional genetically similar Japanese isolates should be sequenced. Recently Vogler described two isolates of *F.t. holarctica* from California of subclade B.Br 002/003 that lie much closer to FSC 022.2 These isolates should also be sequenced if possible. Subspecies *mediasiatica* is represented by only a single genome so genetic diversity is largely undetermined for this subspecies. This is a significant gap as the closest genetic subspecies to *F.t. tularensis* is *mediasiatica*. This gap underscores the importance of sequencing additional isolates of *F.t. mediasiatica*.

Within the *F.t. tularensis* subspecies (type A) the more recently identified[2] divisions between type A I (Shu 4 like) and type AII (ATCC 6233 like) can clearly be seen in the tree. *Francisella novicida* can be seen as genetically distinct from the *F.t.* subspecies of *tularensis*, *mediasiatica* and *holarctica*. Isolates of this *Francisella* species are also generally not pathogenic to normal healthy individuals.[5] However given that *F. novicida* has been isolated from geographically divergent sites in the USA (including environmental aerosol sampling locations) and that a human pathogenic isolate of *F. novicida* has been isolated in Australia and shown to be pathogenic to humans, much more remains to be learned about the genetic and geographic distribution as well as propensity for human pathogenesis of this *Ft.tularensis* near neighbor. It is prudent to sequence additional Australian isolates of *novicida* as well as other *novicida* isolates associated with human disease.

There appears to be less known about the virulence mechanisms of *F.t.tularensis* than of the rest of the Category A bacterial agents. Analyses performed in 2008-9 by LLNL for NBACC (unpublished) have demonstrated that most of the genes that differentiate *F.t. tularensis* from *holarctica* and *novicida* are hypotheticals of unknown function. Additional sequencing of isolates should help to further reduce the set of genes that need to be confirmed via knockouts or other methods to be essential to virulence.

Gaps in sequence data

Virulence & Resistance

While *Franciscella* has been known since the 1920s only in the recent light of genomic sequence data are mechanisms of virulence beginning to be uncovered.[8] In fact in the last three years approximately twenty new publications have described new virulence factors in *F.t.* Of the subspecies *F.t. tularensis* is known to be the most virulent, *holarctica* is less virulent, and due to the scarcity of isolates the human pathogenicity of *mediasiatica* is undetermined. *Franciscella* (species or subspecies) *novicida* was thought to not be pathogenic to humans but a recent isolate from Australia was associated with human disease. Recently a potential component of antibiotic resistance for *Franciscella* has been identified as the AcrAB RND efflux system.[9]

Geolocation

Additional isolates of *F.t. holarctica* from Japan similar to FSC 022 should be sequenced as the current available genome places it closer to *F.t.* than any other current *holarctica* genome. Additional isolates of *F. novicida* from Australia should be sequenced as one has been shown to be pathogenic to humans as well as being genetically diverse.

Additional Species

Since *Francisella* infects over 150 different species of animals, isolates from additional animal species should be considered for sequencing to indicate both essential virulence genes as well as to help determine host range adaptation differences.

Phylogeny

Additional *mediasiatica* isolates should be sequenced as they are thought to be the closest genetic relatives of *F.t. tularensis*. Two recently described[2] California isolates of *F.t. holarctica* subspecies of subclade B.Br 002/003 that are genetically closely related to FSC 022 (a *holarctica* subspecies that is in turn a close genetic relative of *F.t.tularensis*) should be sequenced.

Genetic Near Neighbors

Isolates similar to *Wolbachia* genomes, tick endosymbionts, and soil samples[3] in Figure 2 should be sequenced. Two genomes isolated from clinical samples of human disease that may constitute new *Francisella* sp. should be sequenced.[4]

Potential general sources of *Francisella* material include:

CDC
NAU (Paul Keim)
USDA NVSL (Matt Erdman, Linda Schlater)
BYU (Rich Robison)
Stanford University (Denise Monack)
NIH (Kate Bosio)
LLNL (Amy Rasley)

REFERENCES

1. Worldwide Genetic Relationships among *Francisella tularensis* Isolates Determined by Multiple-Locus Variable-Number Tandem Repeat Analysis. Johansson, A. et Al., 2004, Journal of Bacteriology, 186:p5808.

2. Phylogeography of *Francisella tularensis*: Global Expansion of a Highly Fit Clone. Vogler A.J.,et. Al., 2009, Journal of Bacteriology,191:p2474.

3. Detection of Diverse New *Francisella*-Like Bacteria in Environmental Samples. Barns,S.M., et Al., 2005, Applied and Environmental Microbiology, 71:p5494.

4. Isolation and Characterization of a Novel *Francisella* sp. from Human Cerebrospinal Fluid and Blood. Kugler, K.J., et Al., 2008, Journal of Clinical Microbiology, 46:p2428

5. Molecular Epidemiology, Evolution, and Ecology of *Francisella*. Keim,P., Johansson, A, and Wagner,D.M., 2007, Annals of the New York Academy of Sciences, 1105:p30.

6. Description of two new plasmids isolated from *Francisella philomiragia* strains and construction of shuttle vectors for the study of *Francisella tularensis*. Le Phive, E., et Al., 2009, Plasmid (Eprint).

7. Direct isolation of *Francisella* spp. from environmental samples. Petersen, J.M., et Al., Letters in applied Microbiology, 48:p663.

8. *Francisella tularensis*: unraveling the secrets of an intracellular pathogen. Oyston P.C.F.,2008, Journal of Medical Microbiology, 57:p921.

9. The AcrAB RND efflux system from the live vaccine strain of Francisella tularensis is a multiple drug efflux system that is required for virulence in mice. Bina X.R., et. Al., 2008, FEMS Microbiology Letters, 279:p226.

Anthrax: *Bacillus anthracis*

Summary

Additional isolates from Africa as well as those isolates from North America that comprise the more recently discovered C group should be sequenced. No strong correlation exists between whole genome sequence diversity and the geolocation for samples isolated from either human or animal hosts (Fig. 2 below). However, when considering the geolocations for the completely sequenced genomes of *Bacillus anthracis* (*B. anthracis*) , ten isolates come from North America, seven from Europe, four from Asia as well as four from Africa, which is believed to be the original site of divergence for *B. anthracis*.[1] The African site of origination is supported by the fact that some of the most genetically diverse strains of *B. anthracis* were isolated from chimpanzees in Cote d'Ivoire and Cameroon. [2] There is limited whole genome sequence data available for the near neighbors of *B. anthracis,* namely *B. mycoides* and *B. weihenstephanensis*. Genotypic differentiation of *B. anthracis* from its genetic near neighbors benefits from sequencing of additional near neighbor genomes. Additional isolates of *B. cereus* obtained from human cases of inhalational anthrax should also be sequenced.

Ecology and epidemiology

With the possible exception of Antarctica, there is no continent that is free of *B. anthracis* due to the movement of humans and livestock throughout the globe. *B. anthracis* is globally distributed but many of the most genetically diverse isolates have been obtained in Africa, resulting in speculation that Africa is the origin of *B. anthracis*. [1, 2] *B. anthracis* is also widespread in South and Central America, the Caribbean, Southern and Eastern Europe, Asia and the Middle East. Although anthrax is primarily a disease of herbivores, all mammals, including humans are susceptible. A review of ProMED posts for human cases contracted in the last three years reveal cases from the USA, India , Australia, Indonesia, Armenia, Kazakhstan, Romania, Russia, Zimbabwe, Peru, Guinea-Bissau, Kyrgyzstan, Afghanistan, Viet Nam, United Kingdom, France, Argentina, and Ghana. The human cases consist of infections of all three forms, inhalational, cutaneous and gastrointestinal.

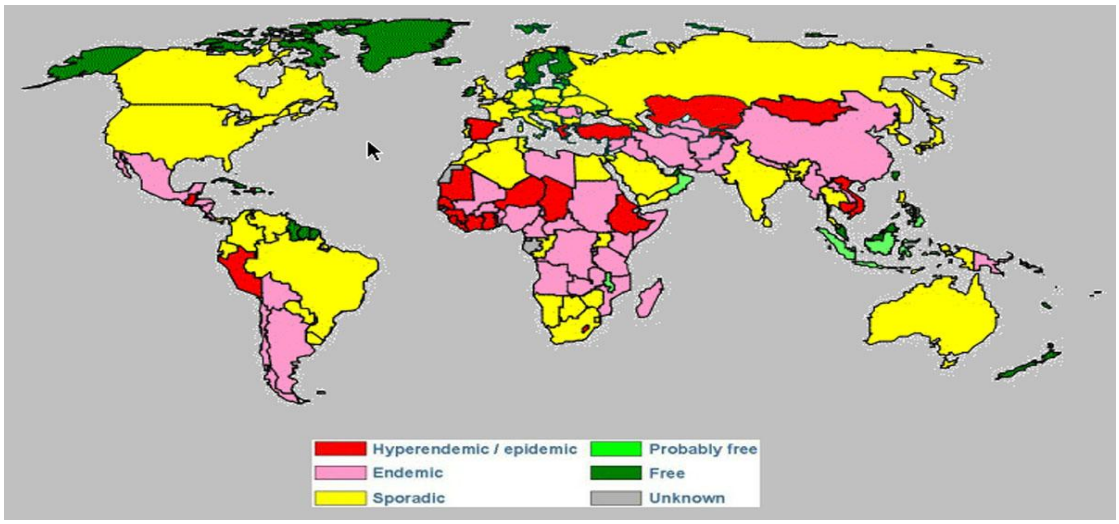


Fig. 1. Global map of *B. anthracis* incidence from the world anthrax data site from Ref. 3.

B. anthracis can be isolated from soil throughout the globe. As shown in Fig. 2 below there is no correlation between the geographic site of isolation and current phylogenetic SNP alignment of the sequences of all of the *B. anthracis* completely sequenced genomes. Since the disease in herbivorous animals is acute, it is usually only possible to isolate the bacteria from animals during an outbreak. No chronic carriers of *B. anthracis* have been identified.

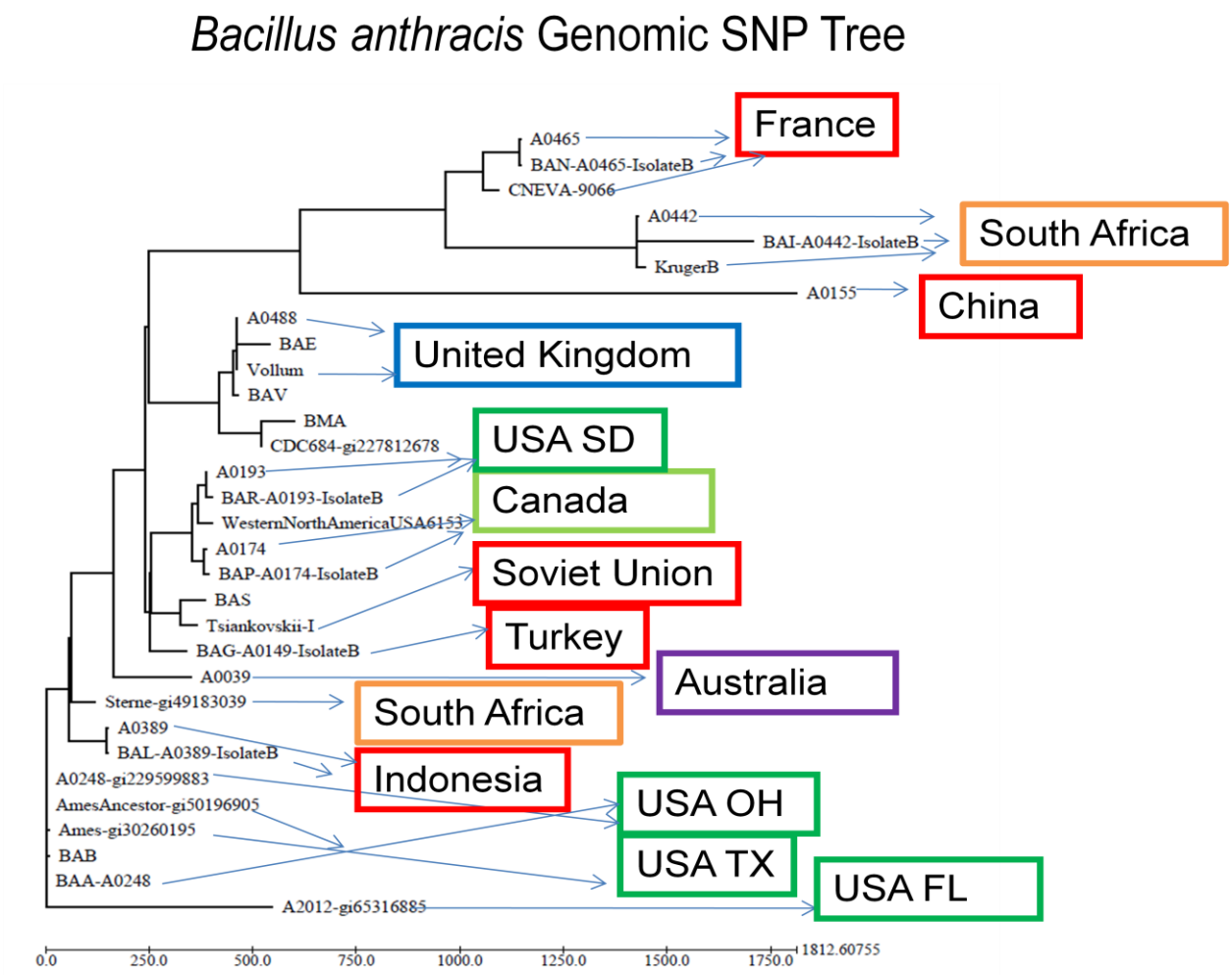


Fig. 2. Phylogenetic tree based on SNP alignment (Gardner et. al unpublished LLNL bioinformatics) of all completely sequenced BA genomes.

Bacterial taxonomy and genome diversity

B. anthracis is believed to be a recently emerged pathogen. The genome of *B. anthracis* is 5.5 Mbp in size and in addition it has two plasmids, pXO1 and PXO2. The *B. anthracis* life cycle which consists of long periods as quiescent spores in the soil, punctuated by rapid proliferation during the short period of bacterial replication prior to death of the host, does not allow opportunity for the genome to accumulate DNA mutations. The result is relatively little genetic variation between *B. anthracis* isolates. However two major but closely related groups of *B. anthracis* strains (A and B) have been described along with a minor more recently identified C group.[4] Group B strains are almost exclusively restricted to southern Africa, while group A is found throughout the world and may have been spread by the domestication and international trade of livestock.[5] The few isolates that comprise the C group come from the southern United States. Interestingly the ancestor of *B. anthracis* is thought to lie between the A and B and more recently identified C groups.[4]

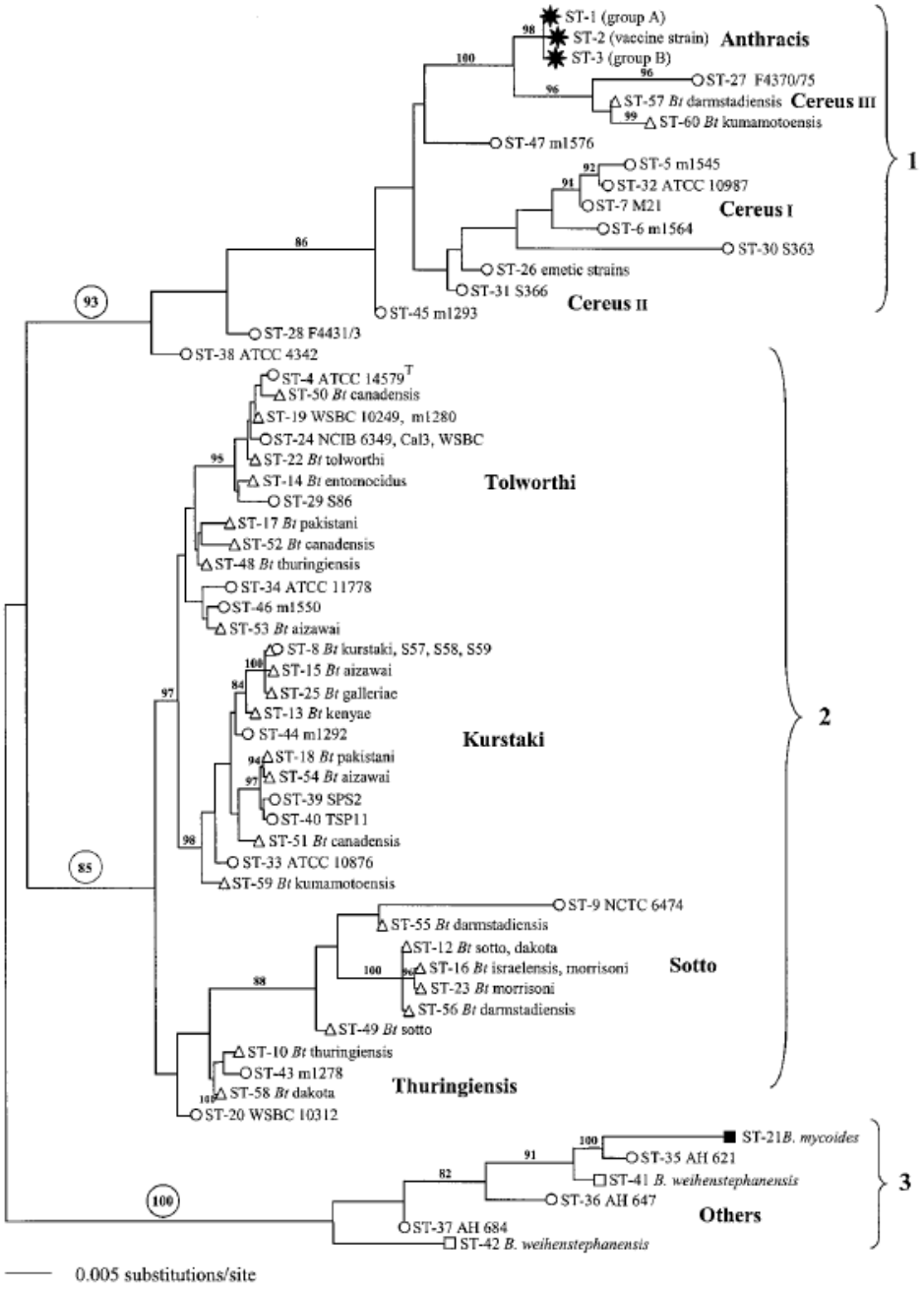


Fig. 4. Maximum Likelihood phylogenetic tree for the concatenated gene sequences for 59 genomes from Ref. 7.
Strain IDs: * *B. anthracis*, ○ *B. cereus*, Δ *B. thuringiensis*, ■ B. *mycoides*, ▯ B. *weihenstephensis*. Clade 1 mainly cereus, 2 mainly kurstaki and thuringensis, Clade 3 genetically distinct isolates are indicated by numbers.

B. anthracis appears to be derived from the clonal expansion of a single ancestral *B. cereus* that acquired the two virulence plasmids (pXO1 and pXO2) and the nonsense *plcR* mutation. Strains of other members of the *B. cereus* group that diverged close to this species boundary are being discovered principally because they share many *B. anthracis*–like traits, but correct nomenclature is dependent on determining where isolates fall in relation to this boundary.[6] The close genetic similarity between all *B. anthracis* isolates when compared to *B. cereus* and *B. thuringiensis* is shown in Fig. 4. In addition it can be seen

that all B. anthracis genomes comprise a single genetically similar group while B. cereus and B. thuringiensis genomes map throughout the phylogenetic tree.

The *Bacillus cereus* species group also includes *Bacillus thuringiensis, mycoides, weihenstephensis* and *cytotoxicus*. The number of completely sequenced genomes is roughly similar between *B. anthracis* (30), cereus (43) and *B. thuringiensis* (27), but drops off rapidly to only three genomes sequenced for B. *mycoides* and one each for *weihenstephensis* and *cytotoxicus*. *B. cereus* causes gastrointestinal distress, necrotic enteritis, liver failure, bacteremia, endocarditis, meningitis, pneumonia and skin lesions.[8,9] Although considered non-pathogenic and used extensively for insect pest control, *B. thuringiensis* has been implicated in burn wound infections and food-poisoning as has *B. cytotoxicus*. Of the other *B. cereus* group members, *B. mycoides*, and B. *weihenstephanensis* only *B*. *mycoides* has been implicated in human eye infections.[10]

Gaps in sequence data

Virulence & Resistance

In a recent review[9] of 666 isolates of *Bacillus* (including *cereus* and *anthracis*), human pathogenic strains were associated with clades 1 and 2 in Figure 2 as were nonpathogenic strains. The pathogenic strains do not have different rates of recombination relative to nonpathogenic strains. There is therefore no strong correlation between a particular genotype and human pathogenicity. A review of the literature reveals no large differences in antibiotic resistance between isolates of *B. anthracis.*

Geolocation

Additional sequence data is needed from Africa. Many isolates from Europe and North America (with the noted exception of the C group) are genetically quite similar suggesting they are all derived from a common ancestor and spread by human migration.[5]

Additional Species

Two *B. anthracis* isolates obtained in 2006 from non human primates in Africa showed increased genetic diversity when compared to all prior isolates.[3] It is recommended, if possible, to sequence additional similar isolates from non human primates.

Phylogeny

The C group of *B. anthracis* has a limited number of isolates and only one completely sequenced genome. Given that the ancestor of B*. anthracis* is thought to lie between the A, B and C groups, [4] it is important to sequence additional isolates that map to this group.

Genetic Near Neighbors

B. anthracis is a member of the *B. cereus* species group of *Bacillus,* and strains of *B. cereus* have been isolated from individuals with symptoms identical to inhalational anthrax.[6] The limited sequence information from these isolates indicates they are genetically more similar to *B. anthracis* than other *B. cereus* strains. Additional sequence data is needed for isolates of *B. cereus* from patients with inhalational anthrax.

Other members of the *B. cereus* species group include *B. thuringiensis* for which there is as much sequence information as for *B. cereus* and *B. anthracis,* and *B. mycoides*, *weihenstephanensis* for which there are very few completely sequenced genomes. As illustrated in Figure 4, isolates of *B. cereus* and *B. thuringiencis* map throughout the tree with some very close to *B. anthracis*. Due to the limited sequence

information, it is not clear if additional isolates of *B. mycoides*, and *weihenstephanesis* could show similar phylogenetic diversity with isolates closely related to *B. anthracis* so it is necessary to sequence additional isolates of these members of the *B. cereus* species group.

Potential general sources of material include:

CDC
NAU (Paul Keim)
USDA NVSL (Matt Erdman, Linda Schlater)
BYU (Rich Robison)
LLNL (Paul Jackson)

REFERENCES

1. Molecular Evolution and Diversity in Bacillus anthracis as Detected by Amplified Fragment Length Polymorphism Markers Keim P., et al., Journal of Bacteriology, 1997, p818.

2. A New *Bacillus anthracis* Found in Wild Chimpanzees and a Gorilla from West and Central Africa Leendertz F.H., et Al., PLoS Pathogens, 2006, 2(1):e8.

3. http://www.vetmed.lsu.edu/whocc/mp_world.htm.

4. Phylogenetic discovery bias in *Bacillus anthracis* using single-nucleotide polymorphisms from whole-genome sequencing Pearson T., et Al., PNAS, 2004, 101(37):p13536.

5. Global Genetic Population Structure of *Bacillus anthracis* VanErt M.N., et Al., PLos ONE, 2007, 2(5):e461.

6. Anthrax, but not *Bacillus anthracis*? Okinaka, R., et Al., PLoS Pathogens, 2006, 2(11):e122.

7. Population Structure and Evolution of the *Bacillus cereus* Group. Priest, F.G., 2004, 160(23):p7959.

8. Characterization of *Bacillus cereus* Isolates Associated with Fatal Pneumonias: Strains Are Closely Related to *Bacillus anthracis* and Harbor *B. anthracis* Virulence Genes Hoffmaster A.R., et Al., Journal of Clinical Microbiology, 2006, 44(9):p3352.

9. Evolution of pathogenicity in the *Bacillus cereus* group Didelot X, Barker M, Falush D, Priest FG, Systematic and Applied Microbiology, 2009, 32:p81.

10. Susceptibility of *Bacillus anthracis, Bacillus cereus, Bacillus mycoides, Bacillus pesudomycoides* and *Bacillus thruingiensis* to 24 antimicrobials using Sensititre automated microbroth dilution and Etest agar gradient diffusion methods. Luna, V.A., et. Al., Journal of Antimicrobial Chemotherapy, 2007, 60:p555.

Plague: *Yersinia pestis*

Summary

Additional samples of *Yersinia pestis* (*Y. pestis*) that are similar to the highly divergent Angola genome shown in Figure 3 should be sequenced. Isolates similar to *Y. pestis* biovars Pestoides and Microtus genomes (Fig. 3) should be sequenced as they are generally avirulent for humans but virulent for mice. These isolates could provide useful animal surrogates for countermeasure development as well as a deeper understanding of mechanisms of host virulence. Given that the site of origin of *Y. pestis* is believed to be north central Africa (Fig. 4) and isolates from all three biovars can be found in China and central Asia, more isolates from these geographic regions should be sequenced.[1] Finally the genetic near neighbors of *Y. pestis* include *Y. pseudotuberculosis* and *Y. enterolitica* however other near neighbors with only a single complete genome sequence such as *Y. frederickensii* (Fig. 2) should be sequenced.

Ecology and epidemiology

With the possible exceptions of Antarctica and Australia, *Y. pestis* has active foci on every continent. Figure 1 shows a map of the globe with countries in red those with known presence of plague in wild reservoir species. ProMed posts for the last three years have reported human infections for all three forms of plague namely bubonic, septicemic, and pneumonic. Countries of origin for the cases included states in the USA (west of 100 W longitude) of AZ, CO, NM, and WY in addition to the countries of Peru, Venezuela, Libya, Tanzania, Uganda, Madagascar, Algeria, Mongolia and China. The occurrences in Africa underscore the fact that plague has been ever increasing on that continent and that it is also thought to be the site of origin for *Y. pestis*. More than 200 species of rodents are the main reservoirs of *Y. pestis* however humans and other animals also serve as hosts with transmission between hosts by fleas. As other members of *Yersinia* genus primarily cause gastrointestinal disease, the transmission of *Y. pestis* also involves gastrointestinal disease in the flea gut where it replicates to high titers and obstructs the G.I .tract causing the flea to starve and therefore bite more frequently and regurgitate bacteria into the bite wounds. *Y. pestis* is further subdivided into three biovars.

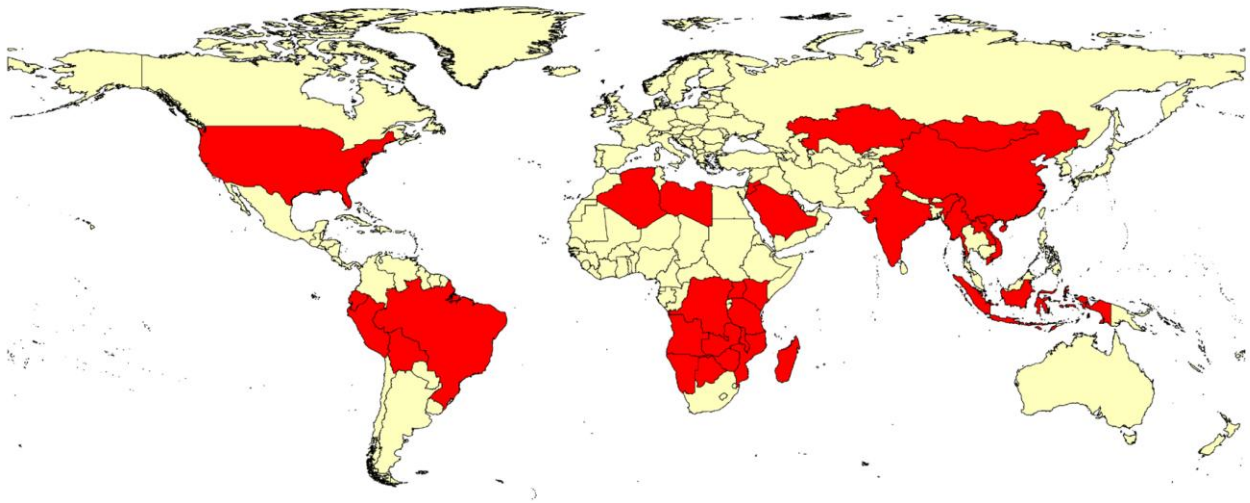


Fig. 1. Global map of *Y. pestis* incidence throughout the world from Ref. 2.

From their current geographic niche and historical records that indicate the geographic origin of the pandemics, researchers have postulated that each biovar caused a specific pandemic. Biovar Antiqua, from east Africa and central Asia, may have descended from bacteria that caused the first pandemic (AD

542), whereas Medievalis, found currently only in central Asia, may have descended from the bacteria that caused the second pandemic (AD 1300). Bacteria linked to the current third pandemic (AD 1894) and all cases of plague isolated from the USA are all of the Orientalis biovar which is currently found worldwide due to human movement and transportation.

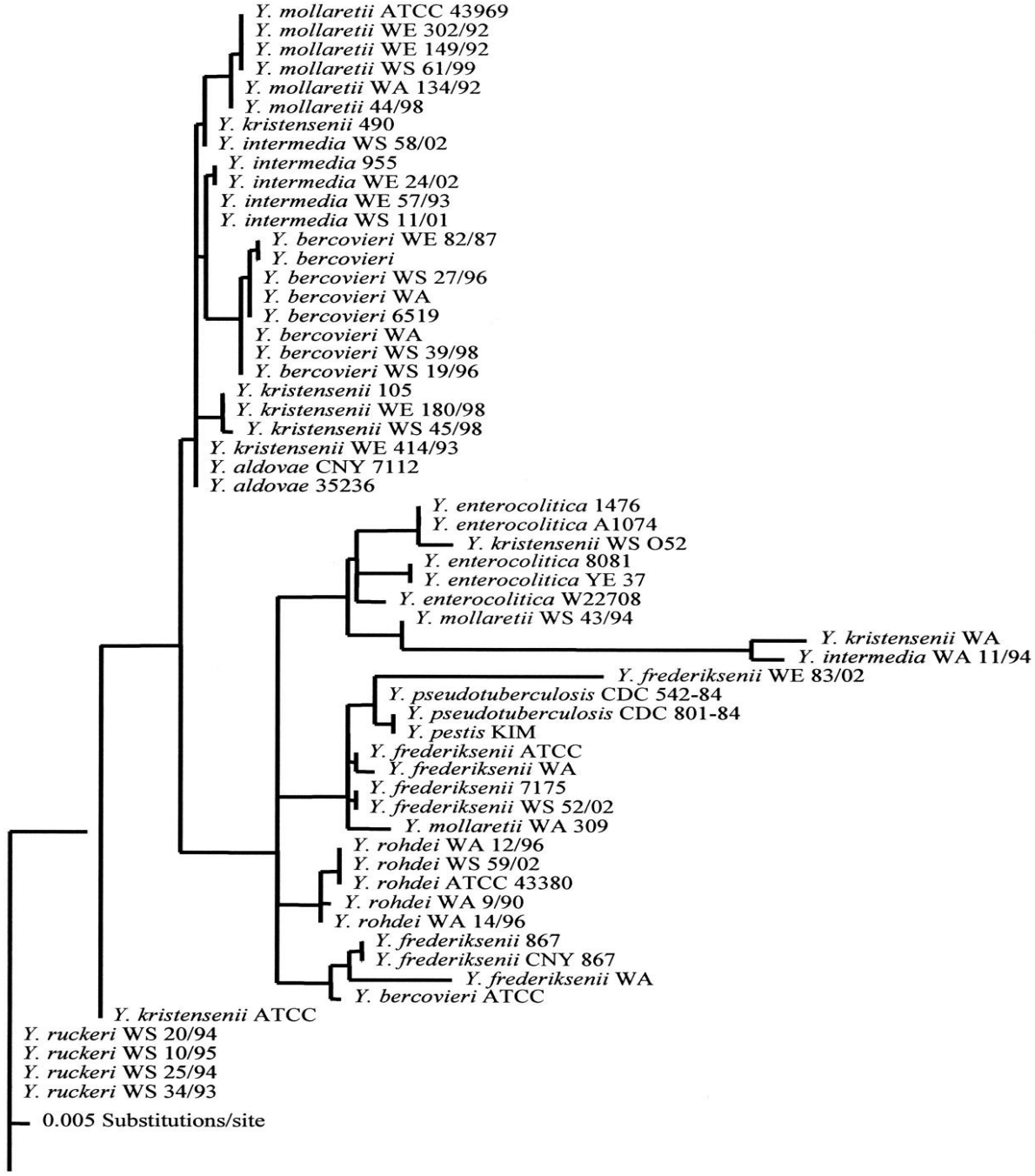


Fig. 2. Minimum likelihood phylogenetic tree of *Yesinia* genus from Ref. 3.

Bacterial taxonomy and genome diversity

The genome of *Y. pestis* is 4.65 Mbp and also contains three plasmids pCD1 (96kbp), pMT 1 (70.3 kbp), pCP1 (9.6 kbp) but only pCP1 and pMT1 are unique to *Y. pestis*. While extensive research and complete genome sequencing (number complete genome sequences in brackets) has been conducted on *Y. pestis*

(25), Y. pseudotuberculosis (5), and *Y. enterocolitica (2),* the remaining eight currently recognized *Yersinia* species (*Y. frederiksenii, Y. intermedia, Y. kristensenii, Y. bercovieri , Y. mollaretii , Y. rohdei , Y. ruckeri ,* and *Y. aldovae*) once collectively called "*Y. enterocolitica*-like" or atypical *Y. enterocolitica* strains have been sparsely researched with only one complete genome sequence of for each of the eight species.[2] The relationship between *Y. pestis* and other members of the *Yersinia* genus are shown in Figure 2.

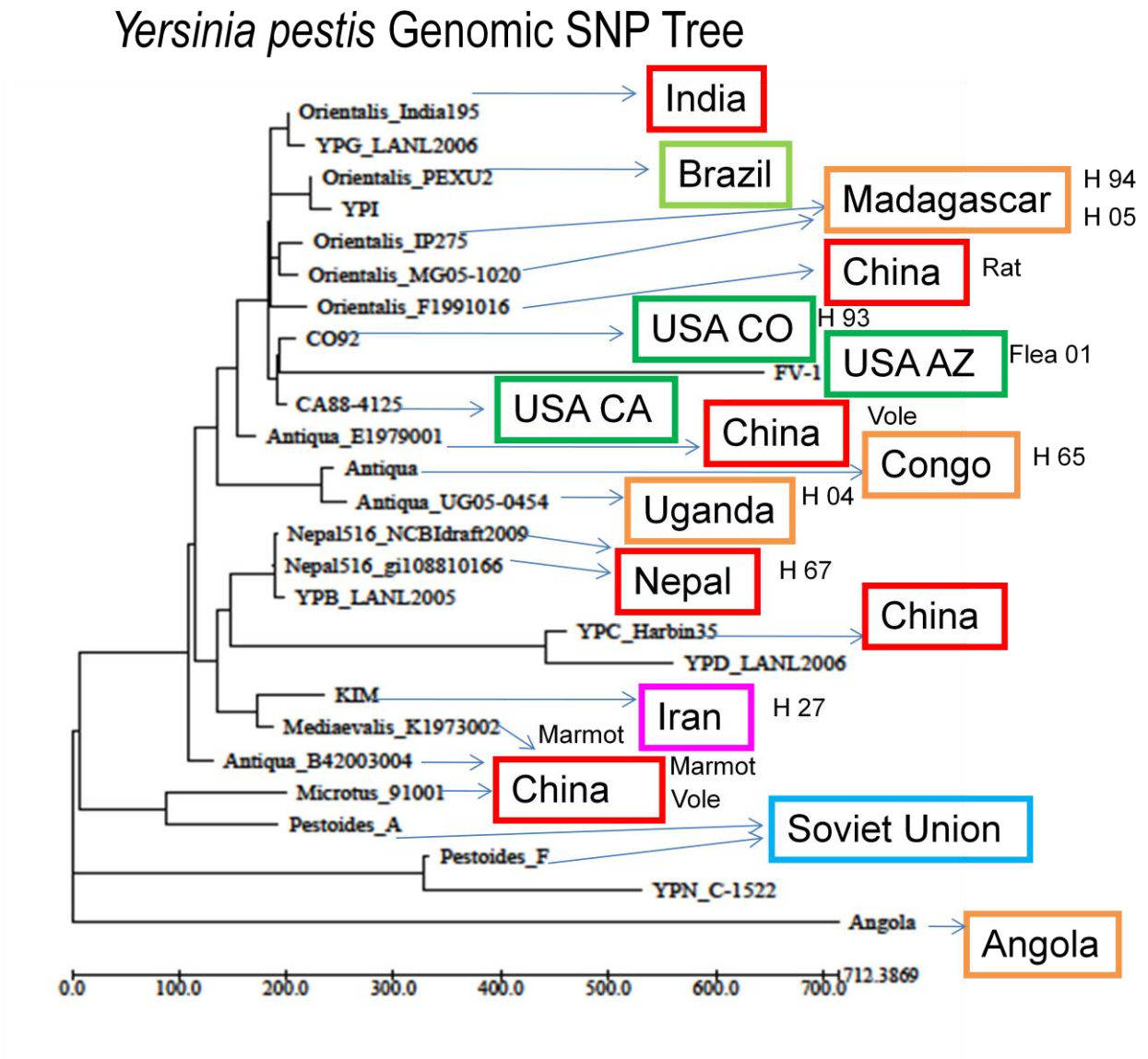


Fig. 3. Phylogenetic SNP alignment of completely sequenced YP genomes. Species of isolation is indicated by an H and the last two digits of year of isolation if human, or the type of animal if non human.
Americas, Europe, Africa, Middle East, Asia.

Y. pestis is believed to have diverged from *Y. pseudotuberculosis* within the last 1,500 to 20,000 years.[4] This divergence was characterized by the acquisition of two plasmids (pMT1, pCP1) that play a key role in flea-borne transmission. In addition to the earlier biovars Antiqua, Mediavalis, and Orientalis a fourth biovar Microtus has been proposed to include strains isolated from rodents that are generally not pathogenic to humans which includes the pestoides strains.[1]

The SNP alignment phylogenetic tree of all the completely sequenced genomes of *Y. pestis* (Fig. 3) shows that a diverse set (with respect to geography and host species) of complete genomes have been sequenced. However *Y. pestis* is found in over 200 species of rodents throughout the globe so it is important to sequence additional isolates of *Y. pestis* similar to Microtus 9001, which was isolated from rodents and is nonvirulent to humans. Microtus-like isolates are most similar to the biovar Antiqua which is believed to be the most ancient and therefore genetically diverse of the three main biovars of *Y. pestis* . The first ten genomes at the top of the tree in Figure 3 are all of the Orientalis biovar. All isolates from the USA cluster together as they are part of this biovar. While this biovar is the most geographically disperse it is also believed to be the most recently diverged biovar of *Y. pestis* and therefore possess the least genetic diversity of the three biovars. At the bottom of the tree is an Antiqua biovar strain Angola which is highly genetically diverse. Therefore additional antiqua isolates from Africa similar to the Angola strain should be sequenced. This approach is also important as the divergence of *Y. pestis* from *Y. pseudotuberculosis* and hence the origin of *Y. pestis* is thought to have occurred in Africa so the most ancient and genetically divergent strains may be found there.[4]

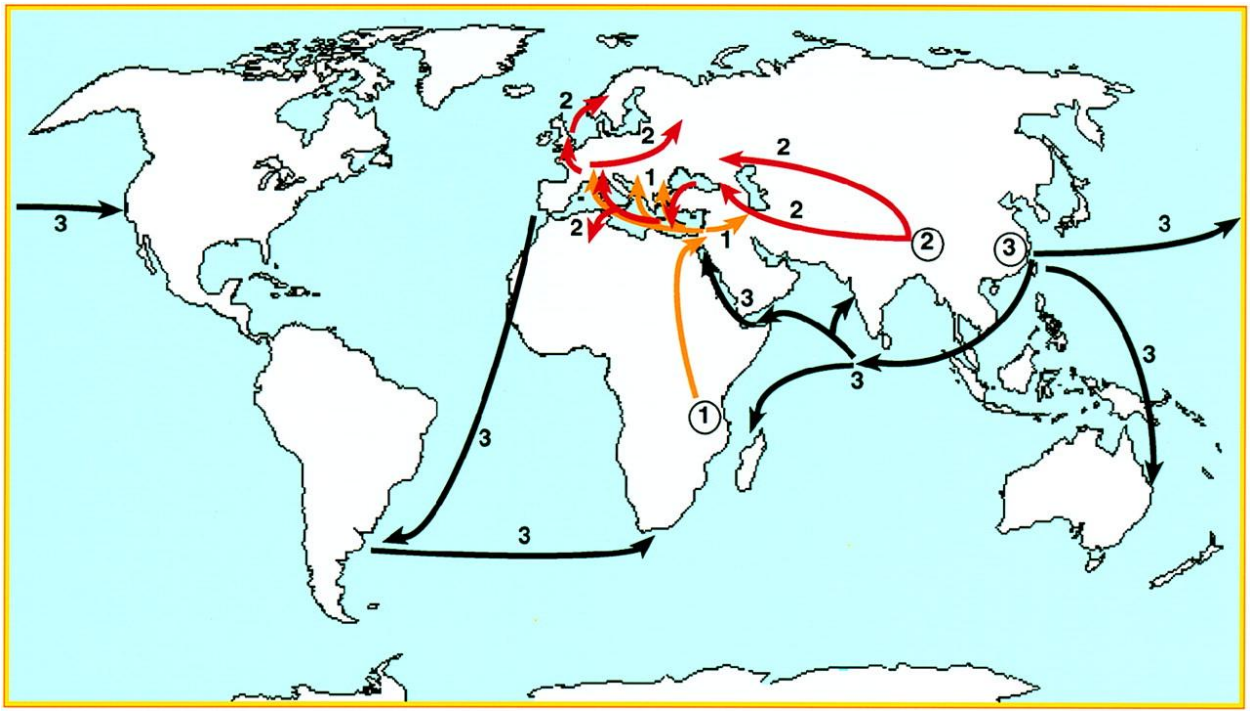


Fig. 4. Global spread and origins of three biovars {antiqua (1), mediavalis (2), orientalis (3)} resulting in three *Y. pestis* pandemics from Ref. 4.

Gaps in sequence data

Virulence & Resistance

Only three species of 15 in the genus *Yersina* are virulent to humans or mammals.[5] The most virulent *Y. pestis* isolates are of the Orientalis biovar. Isolates that are genetically similar to Pestoides and Microtus (Fig. 3) are not virulent to humas while still many are virulent in mice. Additional isolates of Pestoides may provide differences that shed light on human virulence in *Y. pestis*. Rare reports exist of isolates of *Y. pestis* that are resistant to antibiotics but widespread antibiotic resistant strains are not known.

Geolocation
Additional isolates from central Asia (former Soviet Union), China, and Africa should be sequenced as these locations are thought to contain the oldest and most genetically diverse strains of *Y. pestis.*[1]

Additional Species
More than 200 species of rodents are the main reservoirs of *Y. pestis* it is important to sequence additional isolates of YP similar to Microtus 9001 which was isolated from rodents and is avirulent to humans. Microtus-like isolates are also of the biovar antiqua which is believed to be the most ancient and therefore genetically diverse of the three biovars of *Y. pestis* . Finally these isolates could provide a safe model for countermeasure development as well as an extended understanding of the host specificity determining genes of *Y. pestis.*

Phylogeny
As can be seen in Figure 3 the Angola genome is the most diverse of the currently sequenced genomes. This is expected in light of the assumption that *Y. pestis* first emerged in Africa. Additional isolates similar to the Angola strain should be sequenced if possible.

Genetic Near Neighbors
Other than *Y. pestis, pseudotuberculosis, enterolitica* the remaining eight currently recognized *Yersinia* species (*Y. frederiksenii, Y. intermedia, Y. kristensenii, Y. bercovieri , Y. mollaretii , Y. rohdei , Y. ruckeri ,* and *Y. aldovae*) once collectively called "Y. enterocolitica-like" or atypical *Y. enterocolitica* strains have been sparsely studied with only one complete genome sequence of for each of the eight species.[2]

Potential sources of material include:

CDC-Ft Collins (Ken Gage)
NAU (Paul Keim)
Midwest Research Institute (Ted Hadfield)
Porton Down (Rick Titball)
BYU (Rich Robison)
WRAIR (Luther Lindler)
UTMB (Vladimir Motin)

REFERENCES

1. Insight into the Microevolution of *Yersinia pestis.* Clustered Regularly Interspaced Short Palindromic Repeats. Cui Y., et al., PLoS ONE, 2008, p:e2652.

2. Plague: Past, Present, and Future. Stenseth N.C., et Al., PLoS Med, 2008, p:e3.

3. Multilocus Sequence Typing for Studying Genetic Relationships among *Yersinia* Species. Kotetishvili M., et Al., Journal of Clincal Microbiology, 2005, p2674.

4. *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis.* Achtman M., et Al., PNAS (USA), 1999, p14043.

5. Molecular and Darwinian Evolution of Virulence in Yersinia Pestis. Zhou D, and Yang, R., Infection and Immunity, 2009, p2242.

Brucellosis: *Brucella* sp.

Summary

Additional isolates of *Brucella melitensis* (*B. melitensis*), *Brucella suis* (*B. suis*) and *Brucella abortus* (*B. abortus*) from humans should be sequenced because they are scarce relative to animal isolates in the set of completely sequenced genomes. Additional human isolates[1] from *Brucella pinipedialis, Brucella ceti,* and *Brucella inopinata*[2] should be sequenced as they are recently discovered human pathogenic species of *Brucella.* Isolates representing biovar 5 of *B. abortus* as well as biovars 1, 2 and 4 of *B. suis* should have one complete genome sequenced. Bacteria of the genus *Ochrobactrum* are the closest genetic near neighbors to *Brucella*. Isolates of *Ochrobactrum anthropi* and *intermedium* that have been collected from infections of normal healthy humans should be sequenced. *Brucella* is found in the Middle East but isolates from this location are entirely lacking from the completely sequenced genomes.

Ecology and epidemiology

Bacteria of the genus *Brucella* represent some of the globe's major zoonotic pathogens responsible for considerable human morbidity and mortality as well as major agricultural economic losses. Although brucellosis has been eradicated from livestock in the North America, much of Northern Europe and Australia brucellosis remains endemic in most other areas of the world. Sheep (*Brucella ovis* and *melitensis*), goats (*Brucella melitensis*), cattle (*Brucella abortus*), and pigs (*Brucella suis*) are all susceptible respectively to different *Brucella* species which vary widely in host tropism and disease presentation. Many wild animals such as reindeer and rabbits (*B. suis*), elk and bison (*B. abortus*), coyotes (*B. canis*), rodents (*B. neotomae*), voles (*B. microti*), seals (*B. pinnipedialis*), and dolphins (*B. ceti*) can also be infected with *Brucella.* The tenth *Brucella* species *B. inopinata* has been documented to cause human infection[2] but has an as yet unidentified animal host.

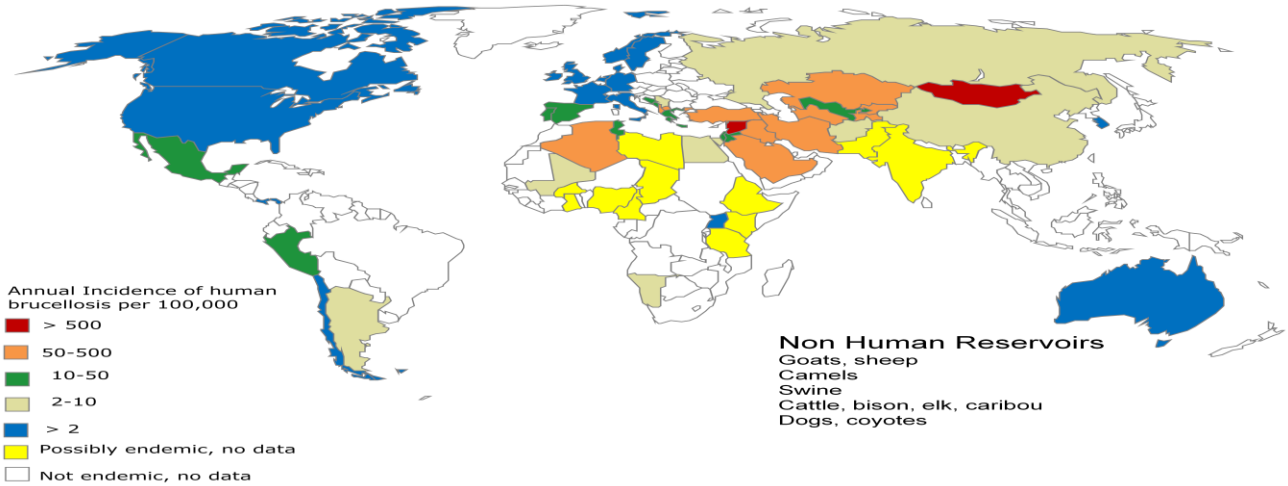


Fig. 1. Global map showing the incidence of human Brucellosis adapted from Ref. 3.

Human brucellosis is the most common zoonotic disease worldwide with an estimated 500,000 cases annually.[3] Virtually all human infections come from direct or indirect exposure to animals and the most

common species responsible for human infection in order of pathogenesis are *B. melitensis*, *suis*, *abortus, canis* (infrequently).

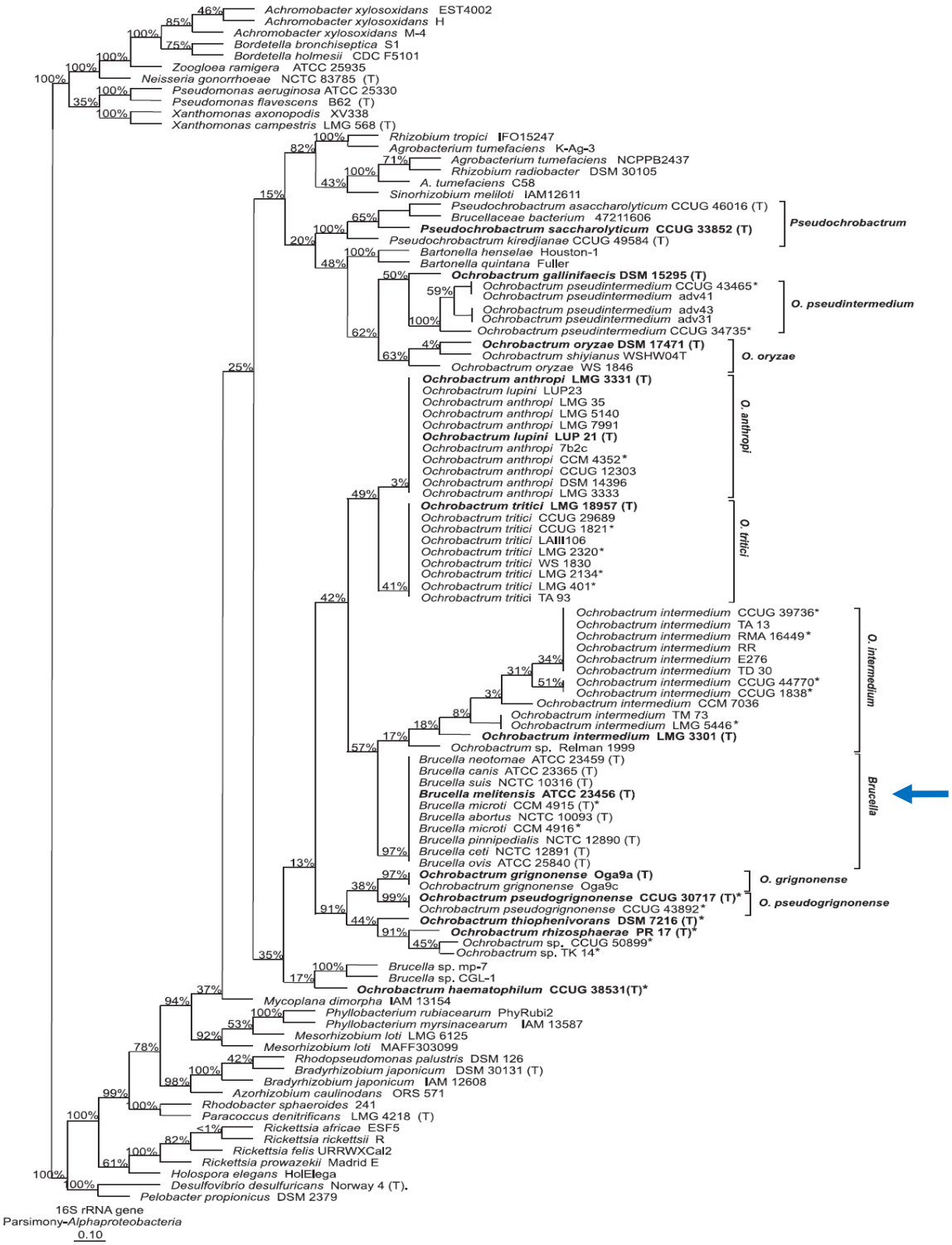


Fig. 2. Phylogenetic position of *Brucella* (blue arrow) in family *Brucellaceae* based on rrs sequence parsimony analysis from Ref. 5.

More recently human infections have also been documented with *B. ceti*.[4] Routes of human infection are most commonly by ingestion of unpasteurized dairy products (goat, sheep and cows milk) but also through cuts in skin, inhalation, or conjuctival contamination.[4] In the absence of treatment infection can lead to chronic disease with debilitating recurring clinical episodes. Effective treatment of humans depends on long periods of antibiotic treatment which often requires combinations of antibiotics. A review of ProMed alerts of Brucellosis documented in the last three years indicates human cases from Kyrgyzstan, China, Bulgaria, Georgia, USA, Greece, Russia, and Mexico.

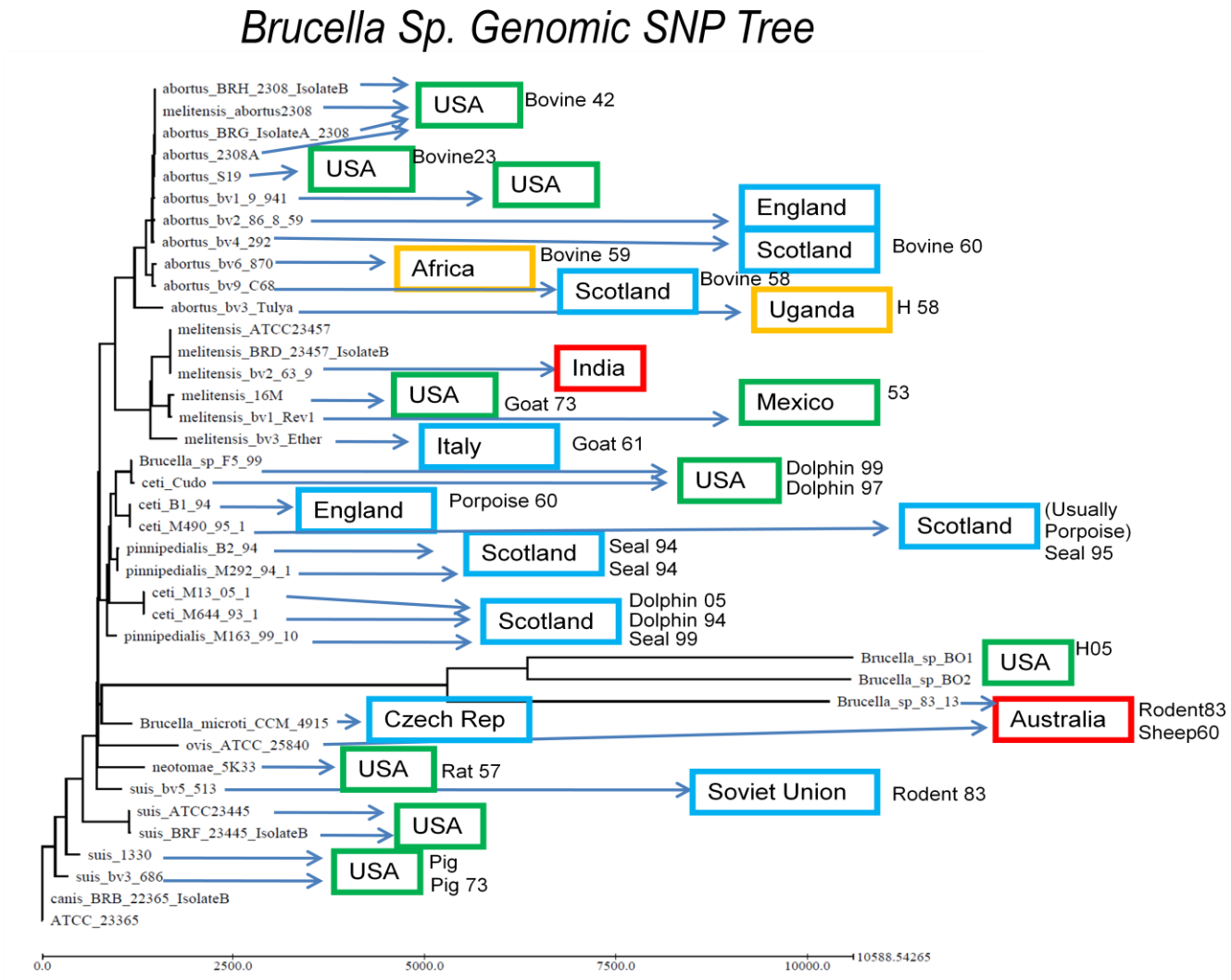


Fig. 3. Phylogenetic tree based on whole genome SNP alignment of completely sequenced *Brucella* genomes (Gardner et. al. unpublished LLNL bioinformatics).
Americas, Europe, Africa, Asia.

Bacterial taxonomy and genome diversity

All *Brucella* species are genetically highly related. In fact DNA hybridization studies have in some cases been used to suggest that all *Brucella* species should be unified as a single species *B. melitensis.* All *Brucella* species have identical 16S and *recA* gene sequences. However the species names have been retained due to biosafety concerns and for practical reasons.[4] In addition to the nine species names of *Brucella* described earlier, species of *Brucella* that have further biovar subdivisions are *B. abortus* (9

however biovar 3 is subdivided into 3a and 3b and 7 is under review and 8 has been removed), *B. melitensis* (3), and *B. suis* (5). *B. suis* biovar three genome has one 3.3 Mbp chromosome while all other genomes of *Brucella* have two chromosomes of 2.1 and 1.2 Mbp. No plasmids of *Brucella* have been identified. Analogous to what is known for *Burkholderia, Brucella* genes on chromosome I have core metabolic functions while those on chromosome II are generally more for adaptation to a variety of environmental conditions. It is also believed that chromosome II may be a megaplasmid that was captured by a *Brucella* ancestor in a similar fashion to *Vibrio cholerae* .[4]

Figure 2 shows the phylogenetic relationship between members of the *Brucellaceae* family and other related bacteria. The position of the *Brucella* species are indicated by the blue arrow in the figure. The closest genetic near neighbors of the family *Brucellaceae* to the *Brucella* are in the genus *Ochrobactrum. O. anthropi* and *intermedium* have been isolated in the past from infections of people with invasive medical interventions however more recently there have been isolates from severe human infections of normal healthy individuals.[6] There is a lack of knowledge concerning the taxonomic position of pathogenic *Ochrobactrum* strains causing severe infections in humans in relation to *Brucella*.[5] There are no completed genome sequences for *O. intermedium* and one for *O. anthropi* so it is important to sequence more genomes of each of these close genetic near neighbors especially isolates associated with human infections.

Fig. 3 is a phylogenetic SNP tree of the completely sequenced genomes that constitute individual *Brucella* species. From top to bottom of the tree they are *abortus, melitensis, ceti, pinnipedialis, inopinata* (BO1), *ovis, neotomae, suis,* and *canis*. The genome of the tenth species *microti* has only recently had a complete genome sequenced.[7] The country of isolation for each genome (if determined) is shown as a colored box. For those isolates for which a date of isolation for the genome was found, the date is listed as a two digit number to the right of the country of origin box. In those cases where a species of origin was found, it is indicated by an H if human or the name of the species is listed between the two digit year and the country of isolation box.

This tree reflects the published relationships between the various *Brucella* species. As can be seen *B. melitensis* has three biovars and is most closely genetically related to *B. abortus.* Biovars 1, 2,3,4,6 and 9 of *B. abortus* are represented by genomes on the tree. As noted earlier *B. abortus* biovar 8 has been suspended and 7 is in dispute as to whether it is a true biovar. A genome of *B. abortus* biovar 5 should be sequenced. For the *B. ceti* species represented on the tree the subdivision between genomes of isolates that infect dolphins and those that infect porpoises can be seen. The new species *B. inopinata* is represented by genome B01. Clustered in this same group is an unusual isolate from a rodent from Queensland Australia. This is interesting as the animal host of *B. inopinata* is as yet unidentified. The next two species *B. ovis* and *B. neotomae* are each represented by one completely sequenced genome. As can be seen biovars 5 and 3 of *B. suis* are represented by one complete genome sequence. However all five biovars of *B. suis* should have at a minimum one completely sequenced genome since it is a human pathogen. Finally the close relationship between *B. canis* to *B. suis* is indicated by its location on the tree. The genome of *B. microti* has only recently been sequenced but it is reported genetically very similar to *B. suis*.7As can be seen there are no completely sequenced genomes isolated from humans or animals from the Middle East. At a minimum isolates of B. melitensis from this region of the globe should be sequenced. Of the 24 genomes for which a species of isolation was identified only two are isolated from humans. Additional human isolates of all human pathogenic strains of *Brucella* (*abortus, melitensis, ceti, inopinata, suis* and *canis*) should be sequenced.

Gaps in sequence data

Virulence and Resistance

Brucella melitensis is the most common human pathogen. However *B. suis, canis, ceti*, and the newly identified *inopinata* have all been isolated from human infections. The differences in virulence (if any) between these human pathogenic species is not understood. [4] Many examples of antibiotic resistance isolates have been described for *Brucella*.[8] However there are a very limited number of completely sequenced *Brucella* genomes isolated from human infections.

Geolocation

Additional isolates of *Brucella* from the Middle East, particularly from humans, should be sequenced. This is important for the geolocation of any future isolates from this region as well as to further define the geodiversity of *Brucella* species.

Additional Species

Additional human isolates of all human pathogenic *Brucella* should be sequenced. In particular additional isolates from newly discovered species *B. inopinata* and from human infections with *B. ceti.*

Phylogeny

B. ceti has unknown genetic complexity comprising at least two subgroups as well as human pathogenic isolates. Additional isolates from *B. ceti* should be sequenced. Only one completely sequenced genome of *B. ovis, neotomae,* and *microti* exists. While these species are not human pathogens more sequencing of these species is needed to fill in gaps in the *Brucella* phylogenetic tree and help to define the genetic determinants of human pathogenicity.

Near Neighbors

The closest genetic near neighbors of *Brucella* the genus *Ochrobactrum* contain species *intermedium* and *anthropi* which have recently been isolated from infections of healthy individuals. Additional genomes from these species should be sequenced.

Potential sources of material include:

CDC (Mike Bowen)
NAU (Paul Keim)
USDA NVSL (Beth Harris, Linda Schlater)
BYU (Rich Robision)

REFERENCES

1. Characterization of a *Brucella* sp. Strain as a Marine-Mammal Type despite Isolation from a Patient with Spinal Osteomyelitis in New Zealand. McDonald. W.L., et al., Journal of Clinical Microbiology. 2006, 44:p4363.

2. Novel *Brucella* Strain BO1 Associated with a Prosthetic Breast Implant Infection. De, B.K., et al., Journal of Clinical Microbiology. 2008, 46:p43.

3. The new global map of human brucellosis. Pappas, G., et al., Lancet Infectious Diseases. 2006, 6:p91.

4. Current understanding of the genetic diversity of *Brucella*, an expanding genus of zoonotic pathogens. Whatmore, A.M., Infection, Genetics and Evolution. 2009, Article in Press.

5. Genetic diversity and phylogenetic relationships of bacteria belonging to the *Ochrobactrum-Brucella* group by *recA* and 16S rRNA gene-based comparative sequence analysis. Scholtz, H.C., et al., Systematic and Applied Microbiology. 2007, 31:p1.

6. Pelvic Abscess Due to *Ochrobactrum anthropi* in an Immunocompetent Host: Case Report and Review of the Literature. Vaidya, S.A., et al., Journal of Clinical of Microbiology. 2006, 44:p1184.

7. Brucella microti: the genome sequence of an emerging pathogen. Audic, S. BMC Genomics. 2009, 10:p352.

8. In vitro antimicrobial susceptibility of Brucella species. Baykam, N., et. Al., International Journal of Antimicrobial Agents. 2004, 23:405.

Appendix I

Rabies species cross-over

In North America, >90% of cases of rabies in animals occur in wildlife, and several mammalian taxa harbor characteristic rabies virus variants (RABVV). In California, striped skunks (*Mephitis mephitis*) and insectivorous bats maintain independent rabies enzootic cycles. Genetically distinct RABVVs have been diagnosed in 12 of the 24 bat species indigenous to California (CA Department of Public Health, unpub. data).While raccoons, coyotes and gray foxes serve as reservoirs for rabies virus in other regions of the United States, striped skunks are the only documented terrestrial mammal reservoir species in California. Typically, interspecies infection with rabies virus produces a single fatal spillover event; secondary transmission has rarely been observed. However, since the end of 2008, a large epizootic of rabies in Northern California has raised concerns not only because there has been a significant spike in attacks by rabid animals on humans and their pets, but also because the primary species involved is gray foxes, and not striped skunks which are the terrestrial reservoir species in this region. Historical surveillance data from the affected area indicate that 0-5 (avg. 1.5) skunks are diagnosed rabid annually. Gray foxes are also occasionally diagnosed with rabies (avg. 1/year). Since October 2008, 4 skunks have been diagnosed with rabies. uring this same time period, 33 gray foxes have tested positive for rabies. Additional foxes exhibiting unusual or aggressive behavior have been euthanized but not tested. It is believed that sustained fox-to-fox transmission is enabling this epizootic to continue and it is hypothesized that molecular changes in the viral genome would be necessary for successful cross-species transmission. While phylogenetic data support that rabies viruses have jumped species boundaries historically, it is rare and has never been subject to comprehensive (full genome) genetic analysis. These epidemiological shifts provide a rare opportunity to study evolution in action to begin to address possible mechanisms and underlying evolutionary forces associated with cross-species transmission.

During recent discussions, Sharon Messenger has indicated that she is willing to work with TMTI to have these skunk and fox samples analyzed to determine the role of genetic drift in interspecies transmission. Sharon has detailed geographical and temporal information for each sample and because these samples are in the form of brain tissue rather than virus isolates generated by infection of cell cultures, sequence

analysis can reflect the true genetic diversity present in nature (i.e. allow analysis of quasispecies using ultra-deep sequence analysis). A white paper is currently being prepared for submission to TMTI.

Appendix II Paul Keim Lab SME input and whitepapers

Questions answered by Dave Wagner and colleagues at NAU, following initial receipt of their whitepapers (which follow these questions). We greatfully acknowledge their gracious assistance.

Bacillus anthracis

What would be your priority for ordering the 3 options? (A, Vollum, Ames/Australia94 branches)
No priority, but I did add a sentence describing what each would accomplish and that we could also take a less in-depth look by doing a bit of all three scenarios.

For each of these options, do you have strains in hand? Accessible via existing collaborations? Or, is sample acquisition required? We have strains – I've made a note about that.

Can you suggest some plausible reasons for TMTI to get excited about these options, besides filling phylogeographical gaps? (e.g., implications for forensics, detection of GE, countermeasure design, etc.). I added some text to suggest that 2 of these scenarios have a direct link to North America.

Do you have any additional isolates of the C group of BA described by Pearson in 2004? I know of only 3 strains in the C group. One is not described in the above paper, but it is being sequenced I think.

Do you have genetic near neighbor strains of BA such as B. Cereus that have been associated with human disease? I only know of a couple, and I think that they have been sequenced. There are also the strains that caused disease in Chimps and Gorillas in the Ivory Coast and Cameroon. These have been sequenced as well, but not yet released.

Do you have isolates of *B. mycoides* and *B. weihenstephensis* that would be helpful to sequence in further refining the phyologenetic tree of *Bacillus*? No we don't have these. We have not been involved in phylogenetic work for the B. cereus group – only the anthracis part of that group, but I do have an interest in defining the phylogeny of the B. cereus near neighbors to Ba.

Yersinia pestis

Very interesting that the Chinese will soon publish on 160 Yp strains from throughout China. How certain are you that they will publicly release the sequence?
They are planning to try and publish in Science so I can't imagine that they would be allowed to publish without releasing the genomes to GenBank. That said, I can't be certain but they have released their genomes in the past and in fact are quite open with their data.

The strains from China shed light on *orientalis* and to some degree *mediavalis* and strains from China are difficult to obtain but Africa is the origin of *antiqua* why do you not want to focus anywhere in Africa except Madagascar given a very diverse genome is from Angola?
It is incorrect that Africa is the origin of Antiqua. The biovars are not good molecular groups, simply phenotypes. The Antiqua phenotype is the ancestral phenotype (various mutations produce the Medievalis and Orientalis phenotypes) and is found throughout the phylogeny of Y. pestis; the Medievalis phenotype is also found in multiple clades throughout the phylogeny. The exception is the

Orientalis phenotype, which corresponds almost 100% to the 1.ORI group. The most basal lineages in the Y. pestis phylogeny are only found in Central Asia, with the exception of the strain called Angola. But no one really knows where this strain actually came from – it just picked up the name Angola somewhere. It is possible that there is a very ancient population of Y. pestis in Africa of which Angola is a member. But it is much more likely that this strain was simply wrongly identified as originating from that country as all of its closest relatives are only found in Central Asia. Think about the Ames strain of anthrax that actually originated in Texas. Given that Madagascar has more human plague cases each year than any other country, this seems a logical place to focus efforts. Plus, we are confident that we can obtain source material from Madagascar. The same cannot be said for other countries in Africa.

Are the 3 projects listed in your priority order?
Yes.

Why should TMTI be excited about sequencing 50-100 North American strains, in particular given that this is the least genetically diverse of the three biovars? (Please give us some ammo to help map this into TMTI's mission needs. How can this aid detection/forensics & countermeasure design?)
As explained in our white paper, the 1.ORI molecular group (again, the scientific community needs to move away from biovars as two of the three are not good genetic groups) is the most geographically widespread of all of the molecular groups within Y. pestis. Thus, potential terrorist around the world are more likely to have access to 1.ORI strains than strains from any other group. Also, strains in this groups are highly virulent and it is unclear that strains from many of the diverse groups in Central Asia are even virulent to humans (see citations in our white paper). The fact that the 1.ORI group is the least genetically diverse is an argument FOR more sequencing in this group, not an argument against it. We need whole genome sequencing within this group to discover rare diversity so it can be used to design assays specific to particular geographic regions. For example, we created an assay designed around a single SNP that can quickly separate all North American Y. pestis from all other global populations. This assay can be used to quickly determine if a bioterrorism event utilized a North American strain, or a non-North American strain. This is a very crucial piece of information in the early stages of a bio-forensics investigation. Most genetic assays available today cannot even separate the 1.ORI group from the other groups, let alone separate 1.ORI strains from Africa from 1.ORI strains from North America. For these reasons, all three of our projects are focused on providing increased resolution within the 1.ORI group. With regards to the focus on North America. A recent report from DHS identified domestic right-wing extremists as a potential source of future terrorist events. Also, plague is endemic throughout the western United States making it relatively easy for a bioterrorist to obtain. Obtaining plague in the United States for use in a bioterrorist event here is probably much more plausible than successfully shipping plague into the United States from another region in the world. Finally, because plague is so common in the western United States it is commonly collected by veterinarians and public health officials. The Amerithrax investigation brought to the forefront the bioterrorism threat presented by rouge scientists and/or medical professional. A detailed genetic landscape of North American plague would be invaluable for a bio-forensic investigation of an event that utilized a North American strain of plague.

Would any of the Madagascar strains be expected to shed light on the increasing antibiotic-resistance encountered there? TMTI could find that very interesting… Do you have access to any strains where unusual A/R is noted?
Probably not. To date, only two resistant Y. pestis strains have been identified – both from Madagascar. One of these strains has already been sequenced and the other has been studied in detail. In both cases resistance was conferred by the acquisition of unique plasmids, probably from some other species.

Project 3 notes that it would provide sequences for all 3 *orientalis* expansions out of China in the 3rd pandemic. Will this provide sufficient data, potentially, to link up with the 160 Chinese genomes and fully establish the evolutionary paths? (Assuming that the Chinese sequence relevant strains, of course.) This could make project 3 more valuable to TMTI than project 1, potentially.
Together, projects 1, 2, and 3 would provide sequences for all three 1.ORI expansions out of China. To do this would require moving forward with all three projects, not just project 3. Note that we have listed a range of the number of sequences that could be generated for each project. So if there is a desire to move forward with all three, we can just reduce the number of sequences for each of the individual projects. It seems likely that the results of these three projects, plus the new 160 genomes from the Chinese, would provide sufficient resolution within Y. pestis to fully establish the major evolutionary paths. Especially considering that the three 1.ORI expansions originated in China.

Do you have any isolates of genetic near neighbors of *Y. pestis* such as *Y. frederickensii* that would be interesting to sequence?
No, and it is not clear how these isolates would be useful to a bio-forensics investigation. It is well established that Y. pestis is a recent clone of Y. pseudotuberculosis so other Yersinia species are not really important from this perspective.

Francisella tularensis
Is your prioritization of the 4 knowledge gaps in the order you list them? Yes.

From TMTI's point of view, knowledge gap 4 might be the most interesting. Do you have the 10-20 strains on hand? Via existing collaboration?
We do not have any of these sorts of near neighbors on hand, neither as isolates nor as DNAs. That is one of the reasons that I changed the text to not list a specific number of strains to be sequenced. I don't know how many we might be able to get a hold of. We are simply aware of their existence from the literature. At this time, we do not have any active collaborations to obtain these sorts of isolates though we can certainly pursue them. This knowledge gap is of interest to us in that it might provide greater insight into the deeper phylogenetic structure of the *Francisella* genus and might aid in the identification of truly *F. tularensis* species-specific markers. That being said, if a *F. tularensis* species-specific marker is defined as only including the three recognized subspecies of *F. tularensis* (*tularensis*, *holarctica*, *mediasiatica* [i.e., NOT including novicida]), then the currently available genomes coupled with extensive screening of near-neighbor DNAs might sufficient for achieving this goal. If a species-specific marker is defined as including subspecies novicida, then additional genomes would definitely be helpful.

Please provide any thoughts you might have to compel TMTI to find Gaps 1 and 3 of great interest to their needs (countermeasure development, better detection/forensics). Do you have the strains you are thinking of sequencing in hand already? Or will they come via existing collaborations?
These subclades represent some of the largest and most genetically diverse subclades that have so far been described. They are also geographically widespread. These characteristics indicate that there is likely significantly more substructure to be discovered. Identifying additional substructure will help in further breaking apart these subclades which would be of use in a forensic investigation (e.g., source attribution). I have added text to the white paper to this effect and changed the numbers of strains to be sequenced to better reflect an overall goal of 50 strains split across multiple knowledge gaps.

For knowledge gap 2 (Iberian Peninsula and France) why should TMTI be excited about filling this gap? Do you have the strains on hand?
F. tularensis subsp. *holarctica* is the most widespread and also the least genetically diverse subspecies within *F. tularensis*. It is also one of the two clinically important subspecies. The lack of genetic diversity makes it very difficult to differentiate among strains, making source attribution difficult for this subspecies. The Iberian Peninsula and France subclade is a particular challenge since it is very widespread and apparently very recently emerged, meaning it has even less genetic diversity than some other *F. tularensis* subsp. *holarctica* subclades. We also have a large set (~100) of French DNAs (not isolates) from this subclade that we can draw on for the study. As Dave said for *Y. pestis*, the lack of genetic diversity is an argument FOR more sequencing since that is likely the only way we will be able to differentiate among isolates from this group. I have added text to the white paper on this. Note that I have reduced the number of proposed strains to 10-12.

Would two isolates from B.Br 002/003 subclades from CA mentioned in the Vogler 2009 publication be interesting to sequence as they define the limits of holartica diversity?
Sequencing one of the isolates from this subclade could be of interest. I have added that as a knowledge gap to the white paper text.

I agree with near neighbor sequencing of *W. persica.* Would tick endosymbionts might be important to sequence and do you have isolates?
As I indicated above, we currently do not have isolates/DNA available for whole genome sequencing for knowledge gap 4, although we could work on obtaining some. With regards to tick endosymbionts, it appears that there are several "Wolbachia endosymbionts" that are sequenced or in the process of being sequenced on GenBank, but they are not classified as species of Wolbachia, rather they are listed as endosymbionts coming from a particular insect species, most of which are not tick species, so I am not sure how useful these sequences would be for understanding the deeper phylogenetic structure of the *Francisella* genus. As I understand it, culturing these endosymbionts is also problematic, which would make obtaining pure DNA difficult.

Burkholderia pseudomallei and B. mallei
You make a great case for this as a source of lateral gene transfer information (this is of key interest to TMTI.) Do you have the isolates on hand for all 3 of the studies listed? Yes, I've added text to that effect.

TMTI's interests may be in opposite priority order from your listing order. Would you try to argue otherwise? If so, please do. I agree and have reversed the order.

For #1 (population dynamics within a single soil sample) are you proposing culturing and isolating the different variants, then sequencing them individually? Or would you try this with a metagenomics approach? A metagenomics approach would be interesting, but may not be informative for learning about recombination. A combination of approaches would be very interesting.

Was Ubon province chosen because it has exceptionally high rates of disease? Or is that a place where you know you will have the least difficulty obtaining samples? Both.

Does this strategy work for *B. mallei* or just *pseudomallei*? Your report is titled *B. pseudo* and *B. mallei* is only mentioned once. Also it seems *B. mallei* is found in central Asia (China and India) not southeast

Asia (Thiland and Australia) like *B. pseudo*. If the strategy is not best for *B. mallei* please add a task/topic for *B. mallei*. I mention mallei only because it is part of the pseudomallei clade. Mallei has a pretty well defined and much simpler phylogeny. There are also very few strains. The amount that we learn from further sequencing of mallei will be a lot less than for any of the other species. Would you like me to draw up a white paper for mallei?

There are only 11 complete genome sequences for *B. mallei* and over twice that number for *B. pseudo*. Do you have access to *B. mallei* strains that warrant sequencing? See above.

Do you know if *B. mallei* is much less prevalent in nature than *B. pseudo*? Yes, it was once globally distributed, but now can only be found in poor regions of the world. Monitoring and sample acquisition in these areas are not what they should be, and getting strains out of those countries is a whole different issue. But we have recently got DNAs from Pakistani recent outbreaks.

Brucella species
For scenario 1 (sequence ~50 isolates from throughout the world) are these isolates currently available to you at NAU or will they need to be obtained via your collaborators?

For scenario 2 (sequence ~50 isolates from the Middle East) also please clarify whether you already have strains from the sampling countries named, have collaborators who you think you can get strains from, or how you would plan to acquire strains. <these 2 questions addressed in updated text>
Is human pathogenicity known for the *ceti/piniped* isolates? Since they are genetically positioned on the tree I sent between *B. melitensis* and B. *suis* (each of which are known to infect humans) perhaps some more *ceti/piniped* isolates should be sequenced?
As for number 3, marine mammal isolates have very rarely been know to infect humans. I only know of a couple of cases. It could be an issue of minimal exposure but I dont think they are really species of medical/biodefense concern. We do have access to samples if you think necessary however.

Population genetics and evolution of *Burkholderia pseudomallei*.

Talima (Tal) Ross Pearson <Talima.Pearson@nau.edu>

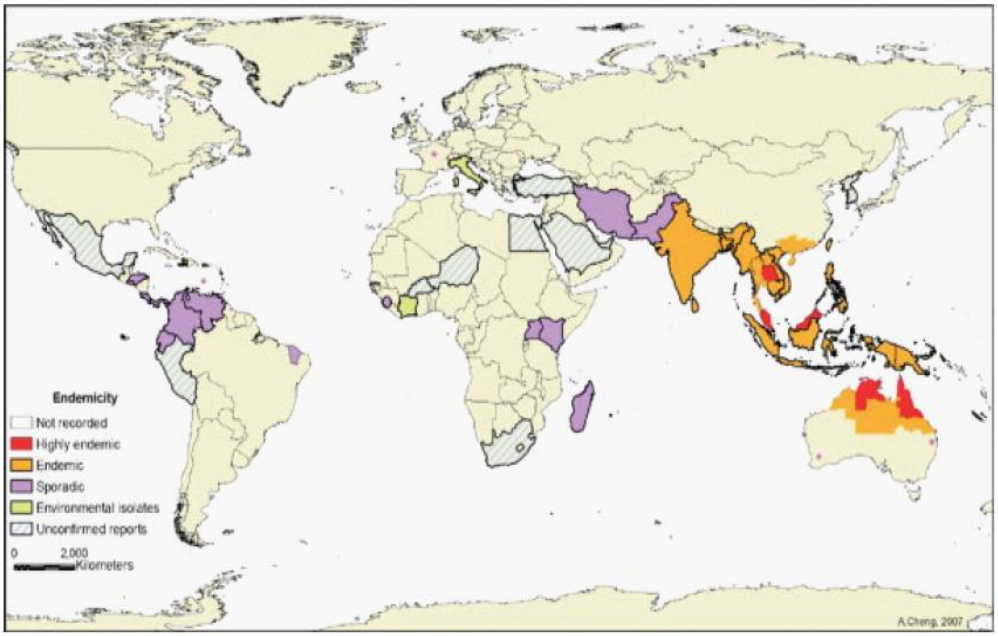
Population genetics and evolution of *Burkholderia pseudomallei*.

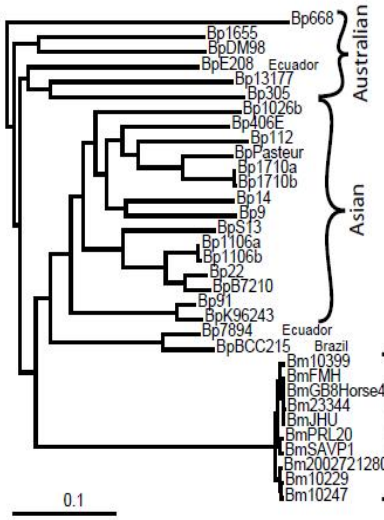
Background and significance: *Burkholderia pseudomallei*, the causative agent of melioidosis is listed by the CDC as a category B biological agent. *B. pseudomallei* is a saprophyte that is endemic to soil and freshwater in the tropical regions of Southeast Asia and Northern Australia. Recent analysis of 33 whole genome sequences suggests that not only are the populations of Southeast Asia and Northern Australia distinct, but also that the Southeast Asian population is a monophyletic derivative from an ancestral Australian population (Pearson and Keim unpublished data). Analysis of MLST data from >1700 isolates and >600 STs suggest that recombination is 18-30 times more likely than mutation to cause allelic changes. With such high rates of lateral gene transfer among lineages, the 7 MLST loci are woefully inadequate for phylogenetic analyses. But, increased genomic sampling through whole genome SNP comparisons can provide necessary phylogenetic resolution. It is therefore apparent that with only 33 whole genome sequences from *B. pseudomallei/mallei*, there are major gaps in our phylogenetic understanding of these species.

The population dynamics of *B. pseudomallei* are highly complex due to the high rates of lateral gene transfer. As expected due to being the ancestral population, the Australian population is more genetically diverse than the monophyletic Southeast Asian population, however the Southeast Asian population appears to recombine more frequently. Detailed information about which regions of the genome are more likely to be subject to lateral transfer, frequency of lateral transfer, and phylogenetic and spatial requirements for lateral gene transfer is unknown. Understanding these parameters will enable us to better reconstruct and understand the patterns of relatedness among this group. As significant knowledge gaps exist at many evolutionary levels, and understanding more at any level will benefit our understanding at other levels, we propose scenarios targeted at understanding the evolution of *B. pseudomallei* at three evolutionary levels. We have strains in our collection for each of these three scenarios.

1) Population dynamics across Thailand or Australia (country scale). From the large MLST database, we can select multiple isolates (25-100) that are the most genetically distant. While this may not correlate completely phylogenetically, we propose selecting the most diverse MLST types from either Thailand or Australia to form the outer bounds of a phylogeny.
2) Population dynamics across Ubon province (regional scale). The province of Ubon in Northeast Thailand has high rates of melioidosis and *B. pseudomallei* is commonly found in the environment. As the relationship of isolates in this region to those found in other regions is unknown, as is the overall diversity of isolates from this region, we propose sequencing multiple isolates (25-100) from across this region.
3) Population dynamics within a single soil sample. In parts of Thailand, there is considerable genetic diversity and population structuring over very small areas [1,2]. MLST and PFGE were used to detect this diversity, therefore it is highly likely that genetic diversity was underestimated as other methods such as whole genome sequencing offer higher genetic resolution. We propose to sequence multiple (25-100) isolates from a single soil sample in order to better understand patterns of relatedness and quantify genetic diversity at a very small spatial scale.

1. Chantratita N, Wuthiekanun V, Limmathurotsakul D, Vesaratchavest M, Thanwisai A, et al. (2008) Genetic Diversity and Microevolution of Burkholderia pseudomallei in the Environment. PLoS Negl Trop Dis 2: e182.
2. Wuthiekanun V, Limmathurotsakul D, Chantratita N, Feil EJ, Day NP, et al. (2009) Burkholderia pseudomallei Is Genetically Diverse in Agricultural Land in Northeast Thailand. PLoS Negl Trop Dis 3: e496.





Francisella tularensis

Amy Joell Vogler <Amy.Vogler@nau.edu>

Background

Francisella tularensis, causative agent of tularemia, consists of three official subspecies: tularensis, holarctica, and mediasiatica. F. novicida, officially another species, is often considered an unofficial fourth subspecies. F. tularensis subsp. tularensis is the most virulent, but is geographically restricted to North America. F. tularensis subsp. holarctica causes a less severe form of disease and occurs throughout the northern hemisphere. F. tularensis subsp. mediasiatica has virulence similar to F. tularensis subsp. holarctica but has only been isolated from a small region in central Asia (5). F. novicida is the least virulent and has only rarely been isolated from North America (5) and once from Australia (7). There are also unculturable genetic near neighbors to F. tularensis that have been identified from the soil (1, 5). Other relatives include tick endosymbionts (e.g., Wolbachia persica) (4, 5) and F. philomiragia (5). However, the deeper phylogenetic structure within this genus is not understood, making the development of species-specific markers problematic.

To date, there are at least 29 complete or draft whole genome sequences available for F. tularensis (Table 1). There are also some F. philomiragia whole genome sequences, although these are so genetically distant as to be relatively useless for identifying SNPs. A recent paper on the phylogeography of F. tularensis used SNPs discovered among eight whole genome sequences to genotype a large number of isolates and identified major subclades in F. tularensis (6). Figures from this paper illustrating the phylogeography for F. tularensis subspecies tularensis and holarctica subclades are presented in Figs. 1 and 2. These figures, which are based upon a canonical SNP analysis of nearly 500 isolates (6), indicate knowledge gaps in our understanding of F. tularensis phylogeography.

Table 1. Current whole genome sequences for F. tularensis.

Sub-clade	Available Whole Genome Sequences
A.I.Br.001/002	FSC 033 (GA, USA), MA00-2987 (MA, USA)
A.I.Br.SCHU S4	SCHU S4 (OH, USA), FSC 198 (Slovakia[a]) , FTN FRAN047, FTB_UNMC061598
A.II.Br.001/002	None
A.II.Br.003/004	None
A.II.Br.004/005	ATCC 6223 (UT, USA)[b]
A.II.Br.006/007	WY96-3418 (WY, USA)[b], FTI WyNC, FTJ LANL2008
B.Br.001/002	FSC 022 (Japan)
B.Br.002/003	None
B.Br.OSU18	OSU18 (OK, USA), MI00-1730 (MI, USA)
B.Br.007/008	None
B.Br.008/009	PTA LANL2006
B.Br.OR96-0246	OR96-0246 (OR, USA)
B.Br.010/011	None
B.Br.FTNF002-00	FTNF002-00 (France or Spain), KO97-1026 (Korea)
B.Br.012/013	None
B.Br.013/014	LVS (Russia)[b], 257 (Russia), RC503 (unknown), FSC 200 (Sweden)
M.Br.FSC 147	FSC 147 (Kazakhstan)
N	U112 (UT, USA), GA99-3548 (LA, USA), GA99-3549 (CA, USA), FTE, FTF M3, FTG, FTH HHS
[a] This genome appears to be a replicate of SCHU S4 (2). It is unlikely that it is truly from Slovakia.	
[b] These strains were placed into the subclade immediately basal to their actual, apparently strain-specific subclade.	

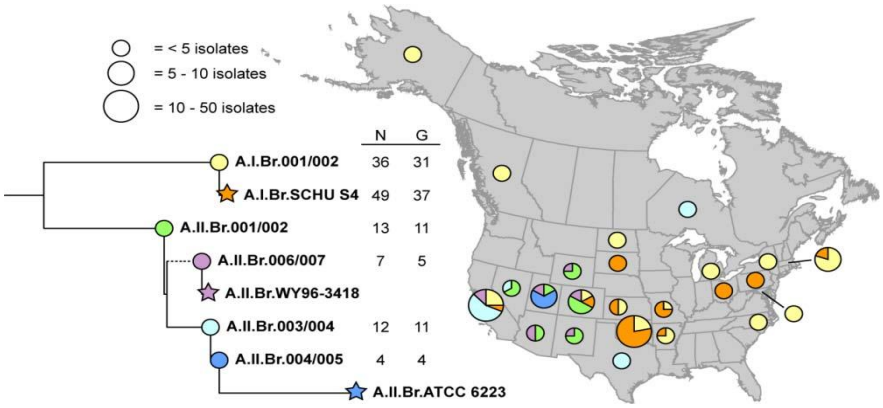


Fig. 1. CanSNP phylogeny and map indicating the frequency and geographic distribution of *F. tularensis* subsp. *tularensis* subclades across North America. The number of isolates (N) and number of MLVA genotypes (G) within each subclade are indicated. Strains belonging in strain-specific subclades (A.II.Br.ATCC 6223 and A.II.Br.WY96-3418) are included within the subclade immediately basal to their actual subclade.

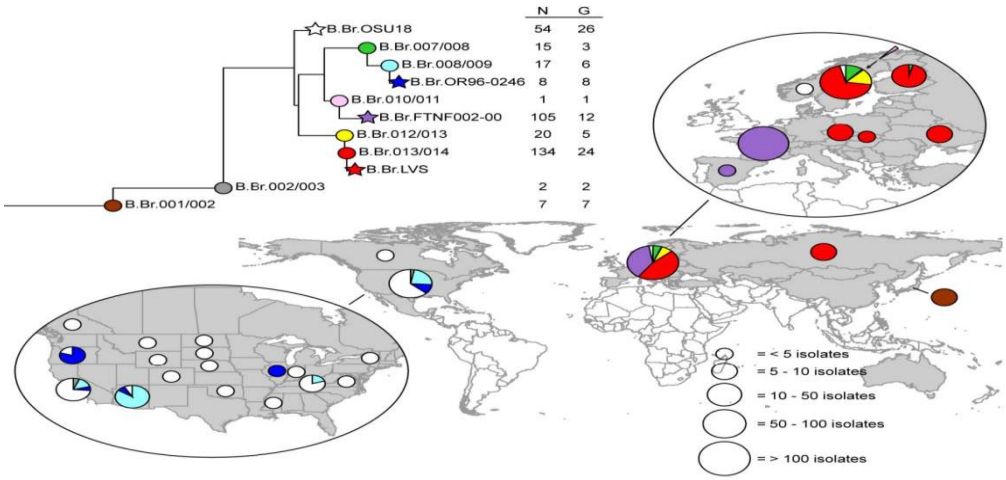


Fig. 2. CanSNP phylogeny and map indicating the frequency and geographic distribution of *F. tularensis* subsp. *holarctica* subclades throughout the world. The numof isolates (N) annumber of MLVA genotypes (G) within each subclade are indicated. The strain belonging in the strain-specific subclade B.Br.LVS is included within thsubclade immediately basaits actual subcB.Br.013/014. Gray regions indicate the known distribution of *F. tularensis*, by country (5).

Knowledge Gaps

F. tularensis subsp. *holarctica* has very limited genetic diversity. A recent phylogeography paper identified 11 subclades and revealed major geographic patterns (6), but several knowledge gaps remain.

Knowledge Gap 1. Subclades B.Br.OSU18 and B.Br.013/014 contain significant MLVA diversity compared to other subclades. In addition these subclades are geographically widespread with B.Br.OSU18 occurring throughout North America and B.Br.013/014 occurring throughout Eurasia (Fig. 2), indicating that these subclades are very good candidates for additional sequencing and SNP discovery. At this time there is one additional genome within B.Br.OSU18 (MI00-1730, Table 1), and three additional genomes within B.Br.013/014 (257, RC503, FSC 200, Table 1) that could be used for additional SNP discovery. Sequencing an additional 10-20 strains from each of these subclades that represented additional MLVA genotypes and/or geographic locations would likely provide the greatest increase in phylogenetic resolution for F. tularensis subsp. holarctica. Increasing phylogenetic resolution in larger subclades such as these would aid in future forensic investigations by providing additional differentiation power and potentially strengthening source attribution capabilities.

Knowledge Gap 2. Subclade B.Br.FTNF002-00 is a very successful clone that has spread throughout France and the Iberian Peninsula and likely elsewhere in Europe (6). There is also very little genetic diversity within this subclade using MLVA (Fig. 2) and as of yet, only one additional genome sequence (KO97-1026, Table 1). Sequencing 10-12 geographically diverse strains from this subclade is probably the only way to identify differences for further differentiating among the members of this very recent and widespread clone. Such differentiation would be critical if a strain from this subclade was implicated in a future forensic investigation.

Knowledge Gap 3. Much remains to be discovered about substructure within *F. tularensis* subsp. *tularensis*. Currently, the only major known phylogeograhic patterns revolve around the genetic and

geographic separation of A.I and A.II (5), but very little is known about any additional phylogeographic patterns within these two groups. Sequencing an additional 10-20 genetically (based upon MLVA) and geographically diverse strains from each of these groups would provide opportunities to further differentiate among isolates within these two major groups and a means of identifying additional phylogeographic patterns, which would aid in future forensic investigations by providing additional differentiation power and potentially strengthening source attribution capabilities for this subspecies.

Knowledge Gap 4. A recent phylogeography paper identified a unique subclade, B.Br.002/003, within *F. tularensis* subsp. *holarctica* that was intermediate between Japanese *F. tularensis* subsp. *holarctica* isolates and typical *F. tularensis* subsp. *holarctica* isolates (6). Sequencing one of the two strains identified as belonging to this unique subclade would increase our knowledge as to the limits of diversity within this subspecies.

Knowledge Gap 5. A previously stated, very little is known concerning the deeper phylogenetic structure within *Francisella*. A whole genome sequence for *W. persica* and any other culturable known genetic near-neighbors to *F. tularensis* such as a recent unusual Spanish isolate from a human infection (FnSp1 (3)) would help to further elucidate the deeper phylogenetic structure within the *Francisella* genus, which, in turn, could help in the identification of species-specific markers for *F. tularensis*.

1. Barns, S. M., C. C. Grow, R. T. Okinaka, P. Keim, and C. R. Kuske. 2005. Detection of diverse new *Francisella*-like bacteria in environmental samples. Appl Environ Microbiol **71:**5494-500.

2. Chaudhuri, R. R., C. P. Ren, L. Desmond, G. A. Vincent, N. J. Silman, J. K. Brehm, M. J. Elmore, M. J. Hudson, M. Forsman, K. E. Isherwood, D. Gurycová, N. P. Minton, R. W. Titball, M. J. Pallen, and R. Vipond. 2007. Genome sequencing shows that European isolates of *Francisella tularensis* subspecies *tularensis* are almost identical to US laboratory strain Schu S4. PLoS ONE **2:**e352.

3. Escudero, R., A. Toledo, H. Gil, K. Kovácsova, M. Rodríguez-Vargas, I. Jado, C. García-Amil, B. Lobo, M. Bhide, and P. Anda. 2008. Molecular method for discrimination between *Francisella tularensis* and *Francisella*-like endosymbionts. J Clin Microbiol **46:**3139-43.

4. Forsman, M., G. Sandstrom, and A. Sjostedt. 1994. Analysis of 16S ribosomal DNA sequences of *Francisella* strains and utilization for determination of the phylogeny of the genus and for identification of strains by PCR. Int J Syst Bacteriol **44:**38-46.

5. Keim, P., A. Johansson, and D. M. Wagner. 2007. Molecular epidemiology, evolution, and ecology of *Francisella*. Ann N Y Acad Sci **1105:**30-66.

6. Vogler, A. J., D. Birdsell, L. B. Price, J. R. Bowers, S. M. Beckstrom-Sternberg, R. K. Auerbach, J. S. Beckstrom-Sternberg, A. Johansson, A. Clare, J. L. Buchhagen, J. M. Petersen, T. Pearson, J. Vaissaire, M. P. Dempsey, P. Foxall, D. M. Engelthaler, D. M. Wagner, and P. Keim. 2009. Phylogeography of *Francisella tularensis*: global expansion of a highly fit clone. J Bacteriol **191:**2474-84.

7. Whipp, M. J., J. M. Davis, G. Lum, J. de Boer, Y. Zhou, S. W. Bearden, J. M. Petersen, M. C. Chu, and G. Hogg. 2003. Characterization of a *novicida*-like subspecies of *Francisella tularensis* isolated in Australia. J Med Microbiol **52:**839-42.

Ancient geographical spread of *Bacillus anthracis*.

Talima (Tal) Ross Pearson <Talima.Pearson@nau.edu>

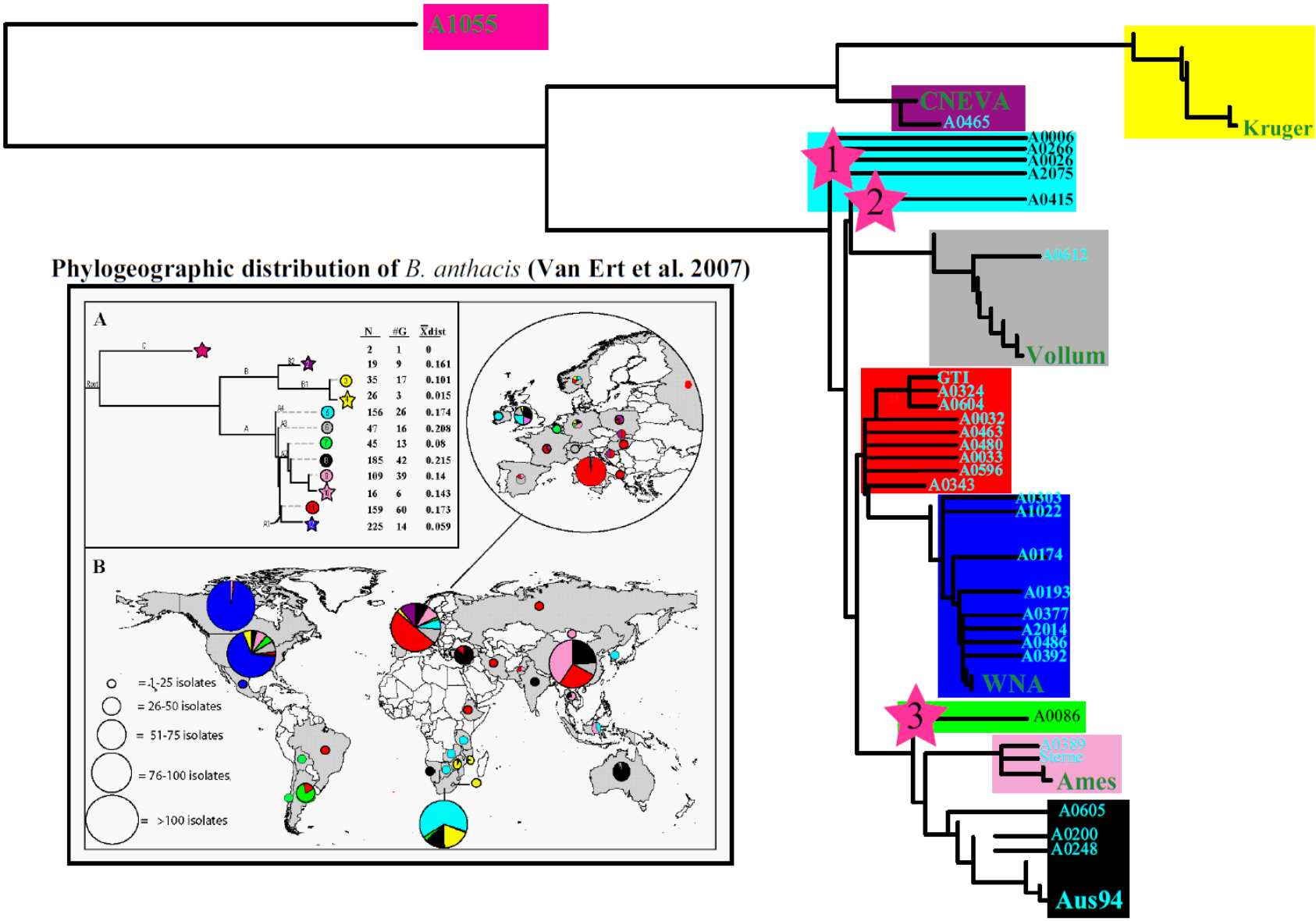
Background and significance: Bacillus anthracis, the causative agent of Anthrax is listed by the CDC as a category A biological agent. B. anthracis also occurs naturally in the soils of many regions in the world, having been dispersed to these regions recently or tens of thousands of years ago. The geographical origin of this species and even the major lineages are unknown, however the accumulation of rare single nucleotide polymorphisms (SNPs) makes it possible to precisely identify evolutionary relationships [1]. Due to the relative rarity of SNPs in the genome, whole genome sequence comparisons are necessary for their discovery. These SNPs can be used to genotype non-sequenced isolates to determine where the evolutionary path of those isolates diverge from the path that connects fully sequenced isolates. However, full phylogenetic topological knowledge of isolates requires whole genome sequencing. As such, whole genome sequencing provides a valuable tool for evolutionary analysis, while sub-genomic genotyping can be used to identify candidates for sequencing to ensure that the choice of strains for whole genome sequencing is sufficient for answering the desired questions.

Members of most B. anthracis lineages can be found in many geographical regions, but their presence is often a result of human trade and commerce over the last century. Source tracking of a given isolate for forensic purposes is most efficient when the geographic region of endemism for that particular lineage is known. For example, members of the Ames lineage have been shipped to labs around the world and can all be traced back to a small ecologically established population in southern Texas [2]. The Texas population, in turn, was probably derived from a Chinese population in the last few centuries [3]. The task of determining the source of more ancient lineages becomes progressively more difficult because the phylogeography of more isolates need to be evaluated. As such, our understand of the Western North American and Ames lineages is substantial [4] and we have recently made progress in extensively genotyping the Australia94 and Vollum lineages. Here, we identify three possible lineages where whole genome sequencing of multiple strains (10-25 strains/scenario) will provide clues for determining the geographical source of these ancient lineages. While the first scenario will fill knowledge gaps regarding the distribution and evolution of anthrax as a whole, the other scenarios will provide particular information on lineages with an important link to North America. Certainly a combination of these scenarios could also be selected. Our collection includes many strains from each of these lineages, so strain acquisition would not be necessary.

Geographical source of the A branch. While much phylogeographic knowledge is known about the A branch, the topology within the first group to diverge from this branch is completely unknown. This group contains 169 isolates collected mostly from southern Africa. As this group is very ancient, we expect to see considerable genetic diversity and expect to identify geographic regions where very little sampling has been done. Phylogenetic knowledge of this group will lead to a substantial increase in our understanding of the genetic diversity of *B. anthracis* (See #1 star on figure).
Geographical source of the Vollum branch. The Vollum lineage was created by the genome sequence of an isolate that is a close relative of strains that were used in BW tests on Gruinard Island, Scotland. But the origins of this lineage may lie in Afghanistan and/or Pakistan from where many isolates were imported to textile mills in the eastern United States (See #2 star on figure).
Geographical source of the Ames/Australia94 branch. It appears likely that these two lineages originated in China, but it is unclear if the common ancestor, shared by these two lineages has a Chinese origin. A lineage to diverge before the Ames/Australia 94 bifurcation point contains 45 isolates, mostly from

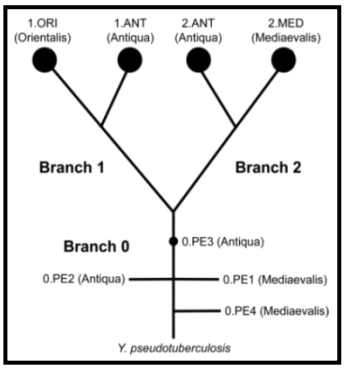
South America. Significant gaps may exist in the geographical disposition of several branches in the extensive collection at NAU and very little is known about the phylogenetic topology of this group.

1. Pearson T, Okinaka RT, Foster JT, Keim P (2009) Phylogenetic understanding of clonal populations in an era of whole genome sequencing. Infect Genet Evol 9: 1010-1019.
2. Kenefic LJ, Pearson T, Okinaka RT, Chung WK, Max T, et al. (2008) Texas isolates closely related to Bacillus anthracis Ames. Emerg Infect Dis 14: 1494-1496.
3. Simonson TS, Okinaka RT, Wang B, Easterday WR, Huynh L, et al. (2009) Bacillus anthracis in China and its relationship to worldwide lineages. BMC Microbiol 9: 71.
4. Kenefic LJ, Pearson T, Okinaka RT, Schupp JM, Wagner DM, et al. (2009) Pre-Columbian Origins for North American Anthrax. PLoS ONE 4: e4813.


A1055
CNEVA
A0465
Kruger
A0006
A0266
A0026
A2075
A0415
1
2
A0612
Vollum
GTI
A0324
A0604
A0032
A0463
A0480
A0033
A0596
A0343
A0503
A1022
A0174
A0193
A0377
A2014
A0486
A0392
WNA
3
A0086
A0389
Sterne
Ames
A0605
A0200
A0248
Aus94
Phylogeographic distribution of B. anthacis (Van Ert et al. 2007)
A
B
N #G X̄dist
2 1 0
19 9 0.161
35 17 0.101
26 3 0.015
156 26 0.174
47 16 0.208
45 13 0.08
185 42 0.215
109 39 0.14
16 6 0.143
159 60 0.173
225 14 0.059
= 1-25 isolates
= 26-50 isolates
= 51-75 isolates
= 76-100 isolates
= >100 isolates

Whole Genome Sequencing Needs in *Yersinia pestis*

David M Wagner <Dave.Wagner@nau.edu>



Background: NAU researchers were part of an international consortium that performed what remains the most up to date published analysis of the global population structure of *Y. pestis* (1). **Fig. 1** is the consensus phylogenetic tree from this study; there are just three major branches (0, 1, and 2) in this phylogenetic tree and eight major molecular groups. This study found that the limited genetic diversity within *Y. pestis* is not evenly distributed across its global range. Most diverse types are still limited to Central Asia as there are specific host association in foci in this region (2, 6) and this local differentiation and adaptation to different hosts and vector species over relatively long periods of time appears to be the main driver of genetic diversity within this species (6). Most global *Y. pestis* populations belong to the 1.ORI group; it is endemic on all continents except Australia and Antarctica. Despite its global distribution, the 1.ORI group is very monomorphic due to a recent genetic bottleneck. This group was responsible for the 3rd pandemic in the 1800s and 1900s, during which it spread from China and was introduced to Africa, Australia, North America, and South America (4).

This same international consortium of researchers has recently completed a new analysis of the global population structure of *Y. pestis* (Achtman et al., unpubl.; **Fig. 2**). In this study we utilized a larger, more diverse strain set and more than 1,200 SNPs. Given the monomorphic nature of *Y. pestis*, this SNP-based study was only possible because whole-genome sequencing approaches were used to discover SNPs.

We feel that new sequencing efforts should be focused on the 1.ORI group. First, this group is the most widespread geographically, including some politically unstable regions. Thus, it seems likely that potential bioterrorist will have easier access to 1.ORI strains compared to strains from other groups. Second, this group is highly monomorphic within a species that is already very monomorphic due to its young age. As such, identifying rare genetic diversity within this group that would be useful for microbial forensics is only really possible by whole genome comparisons. Third, material is very difficult to obtain from Central Asia and it has been suggested that some strains from this region may not even be virulent to humans (2, 3, 5), unlike strains from the highly-virulent 1.ORI group. Fourth, Chinese researchers recently sequenced 160 diverse *Y. pestis* strains from throughout China (Mark Achtman, pers. com.) and are currently writing up their results. Once published, these whole genome sequences will be made available to the wider scientific community.

Proposed Project #1. A detailed phylogeographic analysis of North American *Y. pestis*. Only three whole genome sequences are currently available for North American *Y. pestis*: strains CO92, FV-1, and CA88. Our group has discovered informative SNPs by comparing these three sequences and has identified informative SNPs. We have screened these SNPs across ~700 diverse North American strains and found that there is still significant diversity to be discovered within some of the resulting nodes (**Fig. 3**). We propose to sequence 50-100 additional genomes within North American *Y. pestis*. The SNPs and other genetic variation discovered among these genomes will provide a very detailed genetic landscape of *Y. pestis* in North America.

Proposed Project #2. A detailed phylogeographic analysis of *Y. pestis* in Madagascar. Madagascar now leads the world in human plague cases. Our recent SNP-based global analysis identified multiple genetic subgroups within this population but most isolates were still assigned to two major clades (see the

1.ORI3 group in **Fig. 2**). Currently, only two whole genome sequences are available – both from the same major clade. We propose to sequence 25-50 additional genomes within Madagascar, selecting strains that maximize genetic and geographic diversity. This will provide a better understanding of the population structure of *Y. pestis* in Madagascar.

Proposed Project #3. A detailed phylogeographic analysis of *Y. pestis* in Brazil. *Y. pestis* was introduced to Brazil during the 2nd 1.ORI wave (1.ORI2 group in **Fig. 2**); *Y. pestis* was introduced to North America and Madagascar during the 1st and 3rd 1.ORI waves, respectively. Through a collaborator, we have access to a large and diverse *Y. pestis* strain collection from Brazil. We propose to sequence 10-20 genomes within Brazil. This will provide increased genetic resolution within Brazil but also provide new whole genome sequences from all three waves of the 1.ORI expansion out of China during the third pandemic. Currently there is just one whole genome sequence available within the 1.OR12 group (F1991016).

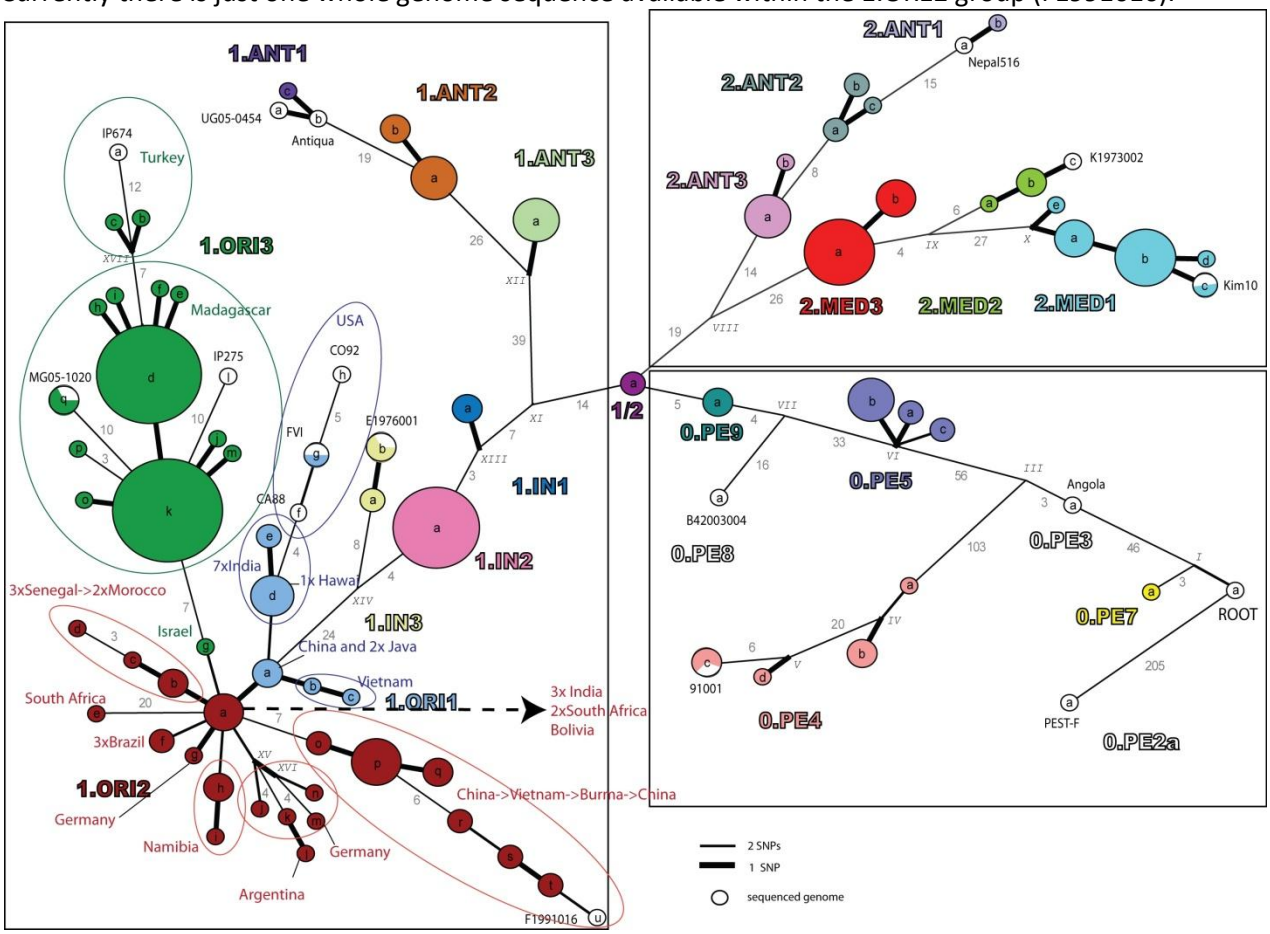


Fig. 2. Updated and expanded SNP-based global phylogeny for Y. pestis. Minimum evolution tree based upon analysis of ~1,200 SNPs screened across a global strain collection (Achtman et al, in prep.). Size of clades (circles) is proportional to number of strains assigned to that clade. Numbers next to branches indicate number of SNPs on that branch. The 1.ORI group spread from China in three waves (1.ORI1, 1.ORI2, 1.ORI3). Most of the structure within this tree is the result of SNPs discovered from whole genome sequences. Clades containing strains that have been whole genome sequenced are presented as white or partially white, with the name of the sequenced strain indicated.

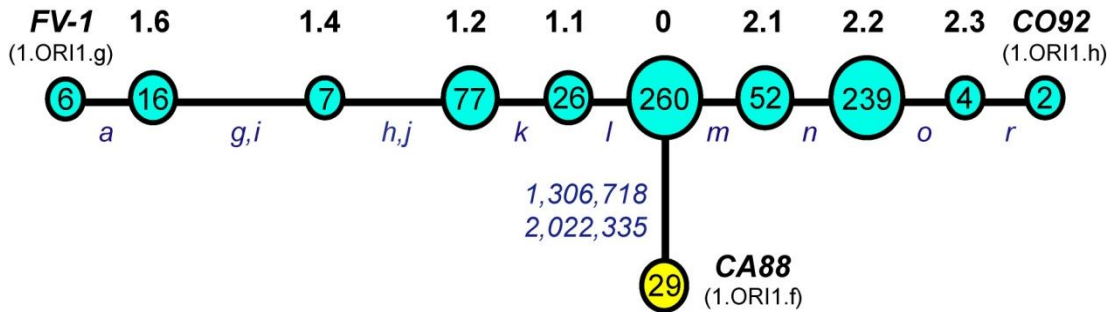


Fig. 3. SNP-based analysis of North American Y. pestis. Phylogenetic tree based upon analysis of 13 SNPs screened across 700 North American isolates of Y. pestis and 18 non-North American isolates. Names of the SNPs, which were discovered by comparing the WGSs of strains CO92, FV-1, and CA88, are listed in blue text below or to the right of the branch on which they fall. The 11 clades identified in this analysis are represented by light blue or yellow circles; the number in each circle refers to the number of isolates assigned to that clade. Clade names are presented above or to the right of the circles. The large number of strains assigned to some of these clades (e.g. clades 1.2, 0, and 2.2) indicates that there is significant undiscovered diversity that could be identified with additional whole genome sequences.

1. Achtman, M., G. Morelli, P. Zhu, T. Wirth, I. Diehl, B. Kusecek, A. J. Vogler, D. M. Wagner, C. J. Allender, W. R. Easterday, V. Chenal-Francisque, P. Worsham, N. R. Thomson, J. Parkhill, L. E. Lindler, E. Carniel, and P. Keim. 2004. Microevolution and history of the plague bacillus, *Yersinia pestis*. Proceedings of the National Academy of Sciences of the United States of America 101:17837-17842.

2. Anisimov, A. P., L. E. Lindler, and G. B. Pier. 2004. Intraspecific Diversity of Yersinia pestis. Clin. Microbiol. Rev. 17:434-464.

3. Fan, Z., Y. Luo, S. Wang, L. Jin, X. Zhou, J. Liu, Y. Zhang, and F. Li. 1995. *Microtus brandti* plague in the Xilin Gol Grassland was inoffensive to humans. Chin. J. Control Endem. Dis. 10:56-57.

4. Perry, R. D., and J. D. Fetherston. 1997. *Yersinia pestis* - Etiologic agent of plague. Clinical Microbiology Reviews 10:35-66.

5. Zhou, D., Y. Han, Y. Song, P. Huang, and R. Yang. 2004. Comparative and evolutionary genomics of Yersinia pestis. Microbes and Infection 6:1226-1234.

6. Zhou, D., Y. Han, Y. Song, Z. Tong, J. Wang, Z. Guo, D. Pei, X. Pang, J. Zhai, M. Li, B. Cui, Z. Qi, L. Jin, R. Dai, Z. Du, J. Bao, X. Zhang, J. Yu, J. Wang, P. Huang, and R. Yang. 2004. DNA Microarray Analysis of Genome Dynamics in Yersinia pestis: Insights into Bacterial Genome Microevolution and Niche Adaptation. J. Bacteriol. 186:5138-5146.

Filling in the gaps in the Worldwide Distribution of Brucella

Jeff Foster <Jeff.Foster@nau.edu>

Background and Justification: Brucellosis is a ubiquitous pathogen of livestock and wildlife and is the most common zoonotic infection worldwide. *Brucella* were the first weaponized bioterror agent and remain a category B Select Agent due to them being highly infectious (fewer than 10 cells), easily grown, and common occurrence throughout the world, particularly in politically unstable regions [1]. Although eradicated from most of Europe and North America, high levels of human infections remain in the Middle East and central Asia.

The genus *Brucella* is currently composed of nine species that appear to have arisen from a non-pathogenic soil bacteria *Ochrobactrum* spp. The *Brucella* research community recently identified the top 30 genomes of research importance and these genomes were sequenced. Sequencing focused on well described biovars and type strains. We have generated a phylogenetic tree based on single nucleotide polymorphisms that are shared among all of the genomes (Fig. 1), following our experience from comparisons of fewer genomes [2,3]. Significant gaps remain in this tree, primarily due to isolate sampling that has been largely restricted to Europe and the US. Whole genome sequencing from undersampled regions is a priority due to the need for increased genetic resolution in this relatively monomorphic genus. We propose focusing on genomes from the three most common and virulent species, *Brucella melitensis*, *B. suis*, and *B. abortus*. These species are the cause of nearly all human infections and represent nearly the complete economic burden for agricultural industries. We present two different scenarios, with either scenario or a combination of the two contributing greatly to our understanding of the distribution and diversity of *Brucella*:

1) Sequencing of 50 geographically diverse isolates from throughout the world, including South America (isolates from Peru and Argentina), Africa (including 1 isolate that is apparent transition between *B. abortus* and *B*. *melitensis*), Asia (including Mongolia, and former Soviet States), Australia (predominantly *B. suis* from feral pigs), and the Middle East (see below). We are collaborating with researchers in England (Adrian Whatmore, OIE Brucellosis Reference Centre, Veterinary Laboratories Agency) and Germany (Holger Scholz, Bundeswehr Institute for Microbiology) who currently possess these samples.

2) Sequencing of 50 isolates exclusively from the Middle East. As the center for livestock domestication, it appears likely that the three focal species originated in this region and spread from there due to commerce. Sampling countries would include Israel, Lebanon, Iraq, Pakistan, Egypt, Saudi Arabia and potentially others. By necessity, we would target areas where isolates could be acquired. The presence of the U.S. military throughout the region increases the potential for sampling from infected soldiers. Drs. Whatmore and Scholz also possess samples from this region and are interested in working with us on this project.

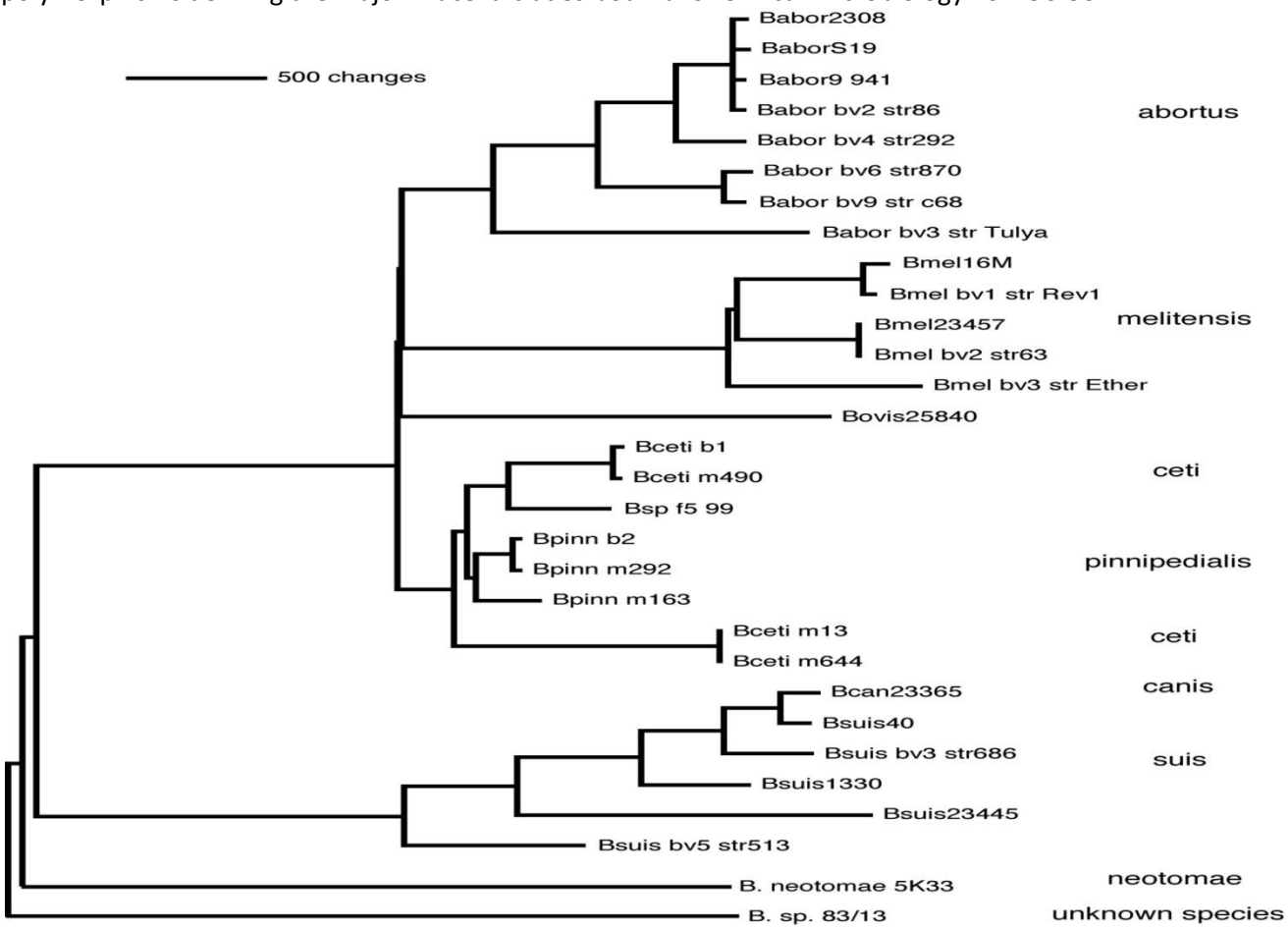
1. Pappas G, Papadimitriou P, Akritidis N, Christou L, Tsianos EV (2006) The new global map of human brucellosis. The Lancet Infectious Diseases 6: 91-99.

2. Foster JT, Beckstrom-Sternberg SM, Pearson T, Beckstrom-Sternberg JS, Chain PSG, et al. (2009) Whole genome-based phylogeny and divergence of the genus *Brucella*. Journal of Bacteriology 191: 2864–2870.

3. Foster JT, Okinaka RT, Svensson R, Shaw K, De BK, et al. (2008) Real-time PCR assays of single-nucleotide

Figure 1. Phylogenetic tree of Brucella genus using 30 genomes from at least seven species. We used only SNPs shared among all of the genomes following Foster et al. 2009.

polymorphisms defining the major *Brucella* clades. Journal of Clinical Microbiology 46: 296-301.



Appendix III USDA NVSL Lab input on sequencing priorities

The USDA National Veterinary Services lab was contacted regarding animal isolates and vaccine strains. Recommendations for *Bacillus anthracis, Burkholderia, Francisella tularensis,* and *Brucella* were gathered from Dr. Matt Erdman and Kristina Lantz and follows below:

Bacillus anthracis

The commonly available live veterinary vaccine strain in the US, *B. anthracis* Sterne strain, is already on one of the trees. Sterne is a pX01+, pX02- excluded strain. The Pasteur strain (BEI NR-38) is an older vaccine strain that does not appear to be on the tree. It is pX01- pX02+ with variable pathogenicity and is a BSL-3 select agent (not excluded). However, it would not be widely available and is not currently used for vaccination. There is an Italian vaccine strain called Carbosap. It is an attenuated pX01+, pX02+ strain. It would be considered a select agent in the US, and is therefore not widely available. Several other vaccine strains that are pX02- have been cited in the literature from other countries, such as Strain 55 in Russia, but these are not commonly found in the US. Currently, excluded strains must be either pX01-, pX02+ or pX01-, pX02-.

The NVSL collection contains 12 virulent strains for which we have relatively complete information on date, area, and species of origin. They range from 1976 to 1999 with 10 bovine and 2 ovine isolates. States of origin are Nebraska, New Mexico, South Dakota, Nevada, and Oklahoma.

Burkholderia mallei* and *B. pseudomallei

There are no avirulent vaccine strains of either *B. mallei* or *B. pseudomallei* available. There are several type strains, formerly from ATCC, that may still be available. For *B. mallei*, ATCC 10399 (BEI NR-21, China 5) ATCC 23344 (BEI NR-23, China 7), and ATCC 15310 (BEI NR-22, Ivan) do not appear to be on trees. For *B. pseudomallei*, ATCC 11668 (BEI NR-24, China 3), ATCC 15682 (BEI NR-25), and ATCC 23343 (BEI NR-26) do not appear to be on the trees. For most of these strains, there is epidemiological information on the BEI website.

The NVSL collection contains 12 field strains of *B. mallei*. There are nine field isolates from Turkey that a researcher brought back here in 1984. They are all Turkish in origin, but the researcher was uncertain of the species or dates of origin. We have a strain of *B. mallei* that was acquired from another lab in 1987, and we believe it is of US origin prior to disease eradication (1930's). We can attempt to get more origin information from the other lab, but unfortunately the researcher involved retired a number of years ago. There are also two field strains from India, one from an equine in 1985, and the other from a mule in 1986. We have three field strains of *B. pseudomallei.* One is from a monkey in North Carolina in 1985, and the other two are from primates in Louisiana in 1993. All three isolates are from research or zoo primates.

Francisella tularensis

Excluded strains of *F. tularensis* include ATCC 6223, LVS (the Live Vaccine Strain), and U112. All three are listed on the tree. Another common strain is Schu4. It is a fully virulent select agent strain often used to make antigen preparations (Schu cells). It is also on the tree.

The NVSL collection contains 20 strains of *F. tularensis* with complete epidemiological information. They range from 1987-2009, and are predominately from rabbits or felines. Only three of the strains are typed to subspecies level; two isolates of *F. tularensis* holartica, one from a beaver in Indiana in 1998, the other from a prairie dog in Ohio in 2001, and an isolate of *F. tularensis* tularensis from a sheep in Wyoming in 1997. There are two isolates from primates in Oregon in 1995 and 1996, and an isolate from a beaver in Illinois in 2005. The remaining isolates are feline or rabbit isolates from Kentucky, Maryland, Wyoming, South Dakota, Pennsylvania, New Jersey, Illinois, Virginia, Ohio, and Idaho.

Recommendations for Brucella from Dr. Beth Harris follow below:

***Brucella abortus* strain RB51 – vaccine strain**

In February 1996, the U.S. Department of Agriculture (USDA) Animal and Plant Health Inspection Services (APHIS) licensed *Brucella abortus* strain RB51 vaccine for use in cattle as part of the cooperative State-Federal Brucellosis Eradication Program. *B. abortus* strain RB51 is a genetically stable, rough morphology mutant that lacks the polysaccharide O-side chains on the surface of the bacteria. These O-side chains are responsible for the development of the diagnostic antibody responses of an animal to brucellosis infection. Therefore, strain RB51 vaccine does not stimulate the production of antibodies on standard diagnostic tests. RB51 has replaced Strain 19 as the recommended vaccine strain in the U.S. Strain 19 vaccine production has ceased, and some States no longer allow vaccination with strain 19. RB51 is also used by various federal brucellosis eradication programs throughout the world.

RB51 was derived by repeated passage of strain 2308 on Trypticase soy agar and varying concentrations rifampin or penicillin, leading to its characteristic rifampin-resistant phenotype. The genetic basis for this phenotype is a single missense mutation in the nucleotide sequence of the *rpoB* gene at codon 526 (Asp →Tyr). The rough phenotype of RB51 is thought to be a result of a spontaneous insertion of IS711 into the *wboA* gene (putative glycosyltransferase gene). However, the extreme attenuation of RB51 compared to other laboratory-derived *wboA* mutants suggests that RB51 carries additional and unknown genomic defects.

Marine mammal *Brucella species*

NVSL has approximately 60 strains of *Brucella sp*. Isolated from marine mammals over the last 10 years. These field strains were recovered from both pinniped (Harbor seal, ringed seal, California sea lion, Northern fur seal, Harp seal and sea otters) and cetacean species (bottlenose dolphin and beluga whale) residing on both coasts of the North American continent. The pinniped species were geographically located in California and Washington on the West Coast, and Conneticut, the Norteastern coast of Canada and Maine on the East Coast. The cetacean species were located in Texas and again the eastern coast of Canada. Although these isolates have not been speciated as either *B. ceti* or *B. pinnipedialis*, they do represent natural infections from wildlife populations in both the Atlantic and Pacific oceans.

Baboon strain of Brucella species.

Also in our collection we have recovered an atypical Brucella sp. from a female baboon with a history of abortion. This animal resided in a primate breeding colony in the southern U.S., but it is unknown if the infection was acquired prior to its residence in the colony. This strain was forwarded to Dr. Adrian Whatmore at VLA, who confirmed it was a member of the Brucella genus. No other characterization of this strain has been done to date.

Appendix IV Access to *Francisella tularensis* clinical isolates with interesting phenotypes

Work by Amy Rasley at LLNL has led to the identification of 7 *Francisella tularensis* proteins they believe may be involved in both environmental persistence and virulence. Their approach to characterizing these proteins has been twofold: 1) generate targeted deletion mutants and screen these mutants in amoeba (looking for those that do not induce encystment) and human monocytes (for those mutants that exhibit an inability to survive intracellularly and 2) clone purify and crystalize each protein. To date, they have screened three of the mutants and have obtained a crystal structure of one of the proteins. Three genes exhibit a phenotype in human monocytes when deleted. The crystal structure they have is from a protein they have named REP24 encoded by FTN_0841 (they hypothesize it is a cysteine protease) and they are close to having quality crystals for a second protein. They are generating the deletion mutant of FTN_0841 as we speak and hope to be able to screen it in both amoeba and human monocytes--this will wrap up their second manuscript on this work.

In addition to the efforts mentioned above, Amy is also in the process of obtaining an *F. tularensis* transposon mutant library and will screen this library in both amoeba and human monocytes. Clones that fail to induce amoeba encystment or fail to survive in human monocytes will be sequenced to identify the gene disrupted.

They have looked at the distribution of these 7 genes across *F. tularensis* subspecies and they have a broad distribution (most are present in all subspecies)--2 of the genes are deleted in LVS which they think may help explain, in part, the inability of LVS to induce amoeba encystment or survive in human monocytes compared with pathogenic *F. tularensis* strains. At this point, they hypothesize that the expression of these genes may be regulated differently in the pathogenic *F. tularensis* strains compared with LVS. In addition, they have sequenced these 7 genes across a panel of pathogenic *F. tularensis* isolates (from Rich Robison's collection) and were hoping to see some deletions, SNPs, etc. that might help us explain the disparate rates of amoeba encystment they observe with the pathogeinc strains. However, there are no major sequence differences. This lends further support to their hypothesis that the genes are being regulated at the level of expression.

Amy recently got results back from the typing that she had done at NAU (Paul Keim's group). All 10 clinical isolates Amy has are Type A which confirms what she already knew using IS element analyses. However, she is also able to distinguish between AI and AII populations. They have 9 AII isolates and 1 AI isolate and the isolates fall into 3 distinct subclades.

Below is a description of the *F. tularensis* isolates. The highlighted rows indicate the isolates Amy thinks we should try to have sequenced by TMTI if possible. They are separated geographically and chronologically are highly virulent in both amoeba and human monocytes.

Recommendations for Sequencing Fully Virulent Type A *Francisella tularensis* Clinical Isolates

ID	Original ID	Alternate ID	Previous Agencies	species	subspecies	Type	Sub-clade
F0680	70102163	Ft 1, reference # 1	unknown	tularensis	tularensis	AII	A.II.Br.001/002
F0681	79101574	Ft 2, reference # 2	unknown	tularensis	tularensis	AII	A.II.Br.001/002
F0682	1365	Ft 3, reference # 3	unknown	tularensis	tularensis	AII	A.II.Br.001/002
F0683	AS1284	Ft 4	unknown	tularensis	tularensis	AII	A.II.Br.001/002
F0684	79400960	Ft 5, reference # 4	unknown	tularensis	tularensis	AII	A.II.Br.004/005
F0685	80700069	Ft 6, reference # 5	unknown	tularensis	tularensis	AI	A.I.Br.001/002
F0686	80502541	Ft 7, reference # 6	unknown	tularensis	tularensis	AII	A.II.Br.001/002
F0687	1385	Ft 8, reference # 7	unknown	tularensis	tularensis	AII	A.II.Br.001/002
F0688	1773a	Ft 9, reference # 8	unknown	tularensis	tularensis	AII	A.II.Br.001/002
F0689	AS2058	Ft 10	unknown	tularensis	tularensis	AII	A.II.Br.001/002