LLNL Genomic Assessment: Viral and Bacterial Sequencing Needs for TMTI, Task 1.4.2 Report

LLNL-TR-423104

January 25, 2010

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
Marisa Lam 925-423-2723, lam9@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.

Task 1.4.2 Report on TMTI sequencing

Under the DTRA Translational Medical Technologies Initiative (TMTI) Lawrence Livermore National Laboratory (LLNL) has been tasked with reviewing sequencing efforts at five sequence service centers. Columbia (Lipkin lab), USAMRIID, and Los Alamos National Laboratory (LANL) are the centers concerned primarily with viral sequencing. Edgewood Chemical and Biological Center (ECBC) and the Naval Medical Research Center (NMRC) are focused on bacterial genome sequencing. The bacteria of interest are the biothreat agents *Bacillus anthracis* (Ba), *Brucella sp.* (Bru.), *Burkholderia mallei* (Burk. m.) and *Burkholderia pseudomallei* (Burk. p.), *Francisella tularensis* (Ft), and *Yersinia pestis* (Yp). The primary viruses of interest are the viral hemorrhagic fever viruses Filoviruses (Ebola and Marburg), and Arena viruses (Guanarito, Junin, Lassa, Machupo, Sabia). Input has been received from the centers regarding what samples are currently being sequenced and this input is summarized below.

Executive Summary

Good progress has been made on both bacterial and viral sequencing by the TMTI centers. While access to appropriate samples is a limiting factor to throughput, excellent progress has been made with respect to getting agreements in place with key sources of relevant materials.

Sharing of sequenced genomes funded by TMTI has been extremely limited to date. The April 2010 exercise should force a resolution to this, but additional managerial pressures may be needed to ensure that rapid sharing of TMTI-funded sequencing occurs, regardless of collaborator constraints concerning ultimate publication(s). Policies to permit TMTI-internal rapid sharing of sequenced genomes should be written into all TMTI agreements with collaborators now being negotiated.

TMTI needs to establish a Web-based system for tracking samples destined for sequencing. This includes metadata on sample origins and contributor, information on sample shipment/receipt, prioritization by TMTI, assignment to one or more sequencing centers (including possible TMTI-sponsored sequencing at a contributor site), and status history of the sample sequencing effort. While this system could be a component of the AFRL system, it is not part of any current development effort.

Policy and standardized procedures are needed to ensure appropriate verification of all TMTI samples prior to the investment in sequencing. PCR, arrays, and classical biochemical tests are examples of potential verification methods. Verification is needed to detect miss-labeled, degraded, mixed or contaminated samples.

Regular QC exercises are needed to ensure that the TMTI-funded centers are meeting all standards for producing quality genomic sequence data.

Sequencing Summary

Existing Samples

Table 1 summarizes the total number of bacterial samples (by genus of bacteria) that have been sequenced as of October and December 2009 and January 2010 under TMTI funding and Table 2 summarizes the viral samples that have been sequenced during at the same time points. Thanks to Greg Meyers for preparation of both of these tables. For the biothreat bacteria 139 *Bacillus*, 5 *Brucella*, 20 *Burkholderia*, 13 *Francisella* and 125 *Yersinia* samples have been sequenced as of January, 2010. Other bacterial samples that have been sequenced are 3 *Acinetobacter*, 2 *Campylobacter*, 2 *Chlamydia*, 8 *Clostridia*, 1 *Escherichia*, 1 *Listeria*, 5 *Rickettsia*, 9 *Salmonella*, and 5 *Vibrio*. A total of 361 bacterial samples have been sequenced.

Organism	Total 01/10/10	Total 12/9/09	Total 10/14
Acinetobacter	3	3	3
Bacillus	139	119	102
Brucella	5	5	5
Burkholderia	20	18	17
Campylobacter	2	2	2
Chlamydia	2	2	
Clostridium	8	2	2
Escherichia	1	1	1
Francisella	13	12	11
Listeria	1	1	1
Rickettsia	5	5	5
Salmonella	9	9	9
Vibrio	5	3	2
Yersinia	125	123	69
Others	23	23	29
TOTALS	361	328	258

Table 1. Number of bacterial samples sequenced as of January 2010.

Of the viral hemorrhagic fevers (in Table 2) 14 Ebola, 27 Marburg, 9 Guanarito, 6 Junin, 15 Lassa, 9 Machupo and 3 Sabia virus samples have been sequenced under TMTI funding. Other viral agents that have been sequenced include 3 Crimean Congo Hemorrhagic Fever (CCHF) viruses, 4 Monkeypox, 6 Rift Valley Fever (RVF), and 1 West Nile (WN) virus. A total of 98 viral samples have been sequenced.

Organism	Total 01/10/10	Total 12/9/09	Total 10/14
CCHF	3	3	3
Ebola	14	13	12
Guanarito	9	9	4
Junin	6	6	4
Lassa	15	15	13
Machupo	9	9	8
Marburg	27	20	16
Monkeypox	4	4	4
RVF	7	6	6
Sabia	3	3	2
WN	1	1	0
TOTALS	98	89	72

Table 2. Number of viral samples sequenced as of January 2010.

ECBC currently has 13 Bacillus samples (both Ba and Ba near neighbors [NN]), 2 Burk. p., 1 Burk. m., no Bru., 6 Ft and 22 Yp samples under TMTI funding in their pipeline. NMRC has 126 Ba and NN, 5 Bru. and NN, 17 Burk. p. and NN, 7 Ft, and 103 Yp and NN.

As for the viruses, we did not have input from Columbia University on the viruses in their sequencing pipeline, as their primary objective is sequencing unknown samples. LANL has 2 Ebola, 3 Marburg, 5 Guanarito, 2 Junin, 2 Lassa, 1 Machupo, 1 Sabia virus samples. USAMRIID has 12 Ebola and 24 Marburg virus samples, 2 Guanarito, 2 Junin, 8 Lassa, 4 Machupo, and 1 Sabia samples being sequenced. Additional viral samples of interest being sequenced at USAMRIID include 1 Crimean Congo Hemorrhagic Fever (CCHF) virus sample, 2 Rift Valley Fever virus (RVF), and 1 West Nile (WN) virus samples.

The bacterial Ba near neighbor (NN) samples being sequenced includes *B. atrophaeus, B. B. cereus, B. coagulans, B. thuringiensis, B. mycoides, B. megatrium, B. pseudomycoides, B. subtilis,* and *B. weihenstephanensis*. Two of these, *B. mycoides* and *B. weihenstephanensis* were identified in the previous 1.4 report as being of particular value for TMTI to sequence. Additional sample metadata is needed from the sequencing centers for us to determine if any of the 31 *B. cereus* samples being sequenced were isolated from cases of inhalational anthrax or if any of the Ba samples come from Africa as recommended in the 1.4 report.

It is encouraging to see one additional Burk. m. sample has been sequenced since there are only 11 complete genome sequences of this bacterial species. Additional sample metadata is needed for us to determine if any of the Burk. p. samples come from non human sources or from Thailand as recommended in 1.4. The Burk. NN being sequenced are *B.oklahomensis, B. thailandensis,* and *B.ubonensis*. The three new samples of *B. thailandensis* will nearly double the number of complete genomes.

The Ft samples being sequenced include 3 *Ft holartica* of which one is a live vaccine strain (LVS) which was a recommendation in the previous1.4 report. In addition *Ft Holartica* samples from Japan, *Ft mediasiatica* samples, *Ft novicida* samples associated with human disease should be sequenced as recommended in the 1.4 report.

The Yp samples being sequenced include one biovar *Yp antigua* sample and *Yp pestoides B, F,* and *G* strains. Pestoides strain samples were recommended to be sequenced as were Yp near neighbors in the 1.4 report. The Yp near neighbors sequenced include *Y. bercovieri*, *Y. mollaretii*, *Y. frederikensii*, *Y. intermedia*, *Y. rohdei*, *Y. ruckeri*, *Y. aldovai*, *Y. kristensenii*, *Y. pseudotuberculosis*, and *Y. enterocolitica* which are all valuable as comparisons to *Yersinia pestis*.

The total number of viral genomes sequenced is approximately one fourth of the number of bacterial genomes sequenced. Progress on sequencing high-threat viruses is being made; however, it is unclear if any of the isolates sequenced are from animal hosts and/or from underrepresented geographic locations. Additionally, no near neighbor viruses to the viral hemorrhagic fever viruses recommended in the previous 1.4 report have been sequenced to date by the sequencing centers. The near neighbor viruses to Lassa are Mopeia, Mobala, and Ippy virus. The Tacaribe virus is a near neighbor of Junin and Machupo viruses. Near neighbors of Guanarito, Sabia and Chapare viruses are Amapari, Cupixi and Tacaribe virus. Near neighbors of Chapare virus are Sabia, Machupo, Cupixi, and Junin viruses. Other relevant near neighbors of all the Arena viruses may remain to be discovered. As pointed out in the 1.4 report, no additional virus isolates may exist for many of these near-neighbor viruses. Discussions with CDC and UTMB should place a high value on any available samples of these poorly-represented viruses.

New Samples for Pipeline

Using information generated in the previous TMTI 1.4 report on sources of bacterial isolates to be sequenced, TMTI has or is pursuing agreements with several parties to provide additional samples for the sequencing pipeline. Northern Arizona University (Paul Keim's lab) has agreed and is already supplying samples to be sequenced by TMTI. The DHS National Biodefense Analysis and Countermeasure Center (NBACC) are now completing an MOU with TMTI to allow some of their samples to be sequenced or provide sequence information to TMTI. They have already provided some samples to ECBC. Unified Culture Collection (UCC) samples have been submitted and TMTI is waiting on a list of samples and a priority list. As of Jan 15, funding mechanisms have been put in place for the UCC and for NAU.

An initial visit was conducted with the Center for Disease Control (CDC) on December 15, 2009 to discuss sharing of sequence information with TMTI. The CDC response was favorable, particularly from *Burkholderia*, *Brucella*, and HFV researchers. They expressed willingness to provide numerous genomes upfront as well as to provide a dynamic, evolving set of samples that

they receive which will serve as a monitor for what is currently circulating. TMTI recognizes that sequence data from these samples is of primary importance and value to the project objectives, and thus is creating an avenue that that could fund CDC researchers to sequence their samples in their own labs and provide the data to TMTI. In late January, extensive input has been received from the CDC indicating samples that could be available for TMTI-sponsored sequencing. We are analyzing this input and will deliver our recommendations separately.

Other potential sources of samples are the Air Force Research Laboratory (AFRL), University of Texas Medical Branch (UTMB) and Brigham Young University (BYU). Tom Brettin of ORNL has also indicated that he may have a path to inquire about access to Battelle samples. (We note that Jim Burans of NBACC has mentioned that some of "his samples" are currently being stored at Battelle; we have not yet determined any potential overlap between this and what Tom Brettin is referring to. Similarly, there are close connections between BYU and both NBACC and NAU that should warn us to be aware of potential sample overlap issues .)

Issues identified

Data and Information Sharing

As of early 2010, data sharing between the TMTI sequencing centers (much less any of the downstream TMTI potential users) has not yet been achieved. The genomes enumerated above have not yet been put into a central system, chiefly because that central system is now being prototyped by CME. We assume that the April 2010 exercise will provide a *de facto* enforcement of sharing, by requiring that all TMTI-funded sequence be in the central database. We also assume that this will be checked and enforced by TMTI management.

Beyond the lack of a proper central repository, there are other cultural and policy factors behind what can only be described as a currently dysfunctional "unified sequencing team". TMTI funds but does not "own" the service lab sequencing centers, and hence appears to be limited in how much it can affect the competitive culture that has been manifested. Access to interesting samples is a critical resource for sequencing centers and there have been examples of missed opportunities to share information with other sequencing centers about who was working on what samples. Assuming that it is actually a goal of TMTI to have a more unified sequencing team, there are some internal team interaction issues that need to be addressed.

There are also significant external issues with regards to data sharing that must be addressed. Keeping the central sequence repository current with the public databases (NCBI, etc.) is a non-trivial problem that presumably CME is now grappling with. A related issue is how to determine when private and public data overlap. This can happen when somebody publicly releases the genome of the same (purportedly) strain as TMTI has done internally. Alternatively, the TMTI center might put a first assembly in the TMTI repository, and then months or years later do a public submission. In the interim, strain names may evolve, as well as the assembly status.

Recognizing and resolving such issues is non-trivial and may affect the quality of downstream countermeasure design (e.g., you'd like the initial draft to be supplanted by the later version for countermeasure design purposes.)

TMTI has launched several efforts to establish quality standards. While this is fine moving forward, there are questions about the consistency of the quality of the work done to date that should be addressed. This includes both within and between the various sequencing centers, and may become clearer once the data has been shared among TMTI participants.

Sample Tracking & Coordination

TMTI is expending a lot of effort (and ultimately, funding) to obtain access to samples that could affect countermeasure design in the near future. It is important that this information be efficiently tracked and coordinated. This could include:

- What locations/individuals have samples of interest to TMTI
- Meta data about such samples (following the MIGS recommendation of Field, 2008, modified for TMTI needs, see reference at end) should include available details on:
    - when and where originally acquired,
    - host species and pathogen species/strain names,
    - why samples are of interest to TMTI (e.g., any unusual phenotype information)
-
- Track which samples have been committed to either being shipped to TMTI or sequenced locally by the strain owner
- Track which samples have been received by the TMTI receiving center, and which genomes have been received (in cases where the sequencing was done locally for TMTI's behalf.)
- Track the disposition of received samples (e.g., which TMTI center(s) are sequencing them)
- Track the status of the samples queued up for sequencing at the TMTI centers. What priority, where in the sequencing process is the sample, etc. Ideally, the local LIMS system could automatically provide such updates.

Such a system could provide TMTI with an efficient cradle-to-grave tracking of strains from the first time they were identified as being potentially useful for TMTI until their sequence is in the TMTI repository. This would be a reasonably complex Web-based database that also would

allow direct program interfaces (from a TMTI sequencing center LIMS, for example.) Note that TMTI would also have to set and enforce policy on usage of such a system.

Sample Verification

Having an initial triage/quality check on incoming samples prior to sequencing is important for saving time and money. This is especially true for archived samples where tubes with hand-written labels are involved. One aspect is to have a multiplex PCR assay to be able to confirm the key pathogens of interest. Another aspect is to have a broad-detection microarray that can additionally detect potential contamination or confirm neighbor species that are not in the PCR multiplex. It is too late to discover problems after sequence has been generated and there are too many things that go wrong to blindly depend on tube labeling as being correct.

Quality Control

It is important to establish a quality control check on the sequence data produced by the 5 TMTI sequencing centers. We support the current proposal to include a QC check as part of each bi-annual exercise, beginning in October, 2010. Initially, this should consist of re-sequencing available strains that have been previously sequenced (complete, not draft) by a non-TMTI center with a community reputation for high standards. Rather than a competition, this should be viewed as a QC learning exercise across TMTI centers who may be using different sequencing technologies/chemistries and potentially different assembly or mapping software. Once baseline quality has been established and maintained across the subsequent exercise, the quality control samples could then be novel (not previously sequenced) samples that the TMTI centers could collaborate on to compare results and derive a consensus answer. The hope is that exercises like these might provide a way for some actual teamwork and collaboration to begin to grow.

Possible Solutions

The issues surrounding data and information sharing need to be addressed by TMTI. This might include team-building exercises, facilitated discussions led by skilled consultants, rotating staff for working visits to other centers, or other standard approaches to build functional teams.

TMTI needs to issue, “sell”, and enforce good policies that will provide all participants with a guided path to do the right things with respect to information and data sharing. Determining what these policies should be may require some work beyond what is already envisioned.

There is a need to have a good tracking system for all work pertaining to the sequencing being done across the multiple TMTI-funded sequencing centers (including potentially CDC and others who may do their TMTI-funded sequencing in-house.) The first step is to have a requirements study followed by a TMTI review. Subsequently, an RFA can be issued for the design and construction of the system. Finally, TMTI will need to establish and enforce policies so that the system is used appropriately by all relevant performers.

TMTI needs to decide how standardized sample verification should be done by their centers. This was discussed in the task 1.1 and 1.2 reports, and action needs to be taken. There are multiple options for multiplex PCR panels (to identify key known agents) and broad-spectrum arrays to identify other agents, including potential chimeras and contaminants. Requirements should be prepared that lead to an RFA for solutions relevant to TMTI-specific needs.

Quality control should be implemented as part of the exercises. It would be a good idea to map out options for QC exercises for the next several years and work these with TMTI management in ways that not only monitor sequencing quality but also increase the chances for the sequencing centers to work together as a true team.

References

The minimum information about a genome sequence (MIGS) specification Field, D. et al. Nature Biotechnology. 2008, 26:p541.