White Paper Application

**Project Title:** Whole genome sequencing of clinical strains of *Mycobacterium tuberculosis*

**Authors:** David Alland, Jerrold Ellner, Susan Dorman, Moses Joloba, Clifton Barry

**Primary Investigator Contact:**

<table>
<thead>
<tr>
<th>Name</th>
<th>David Alland M.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position</td>
<td>Director, Center for Emerging and Re-Emerging Pathogens</td>
</tr>
<tr>
<td>Institution</td>
<td>New Jersey Medical School</td>
</tr>
<tr>
<td>Address</td>
<td>Newark</td>
</tr>
<tr>
<td>State</td>
<td>NJ</td>
</tr>
<tr>
<td>ZIP Code</td>
<td>07103</td>
</tr>
<tr>
<td>Telephone</td>
<td>973-972-2179</td>
</tr>
<tr>
<td>Fax</td>
<td></td>
</tr>
<tr>
<td>E-Mail</td>
<td><a href="mailto:allandda@umdnj.edu">allandda@umdnj.edu</a></td>
</tr>
</tbody>
</table>

1. **Executive Summary** *(Please limit to 500 words.)*

Provide an executive summary of the proposal.

The NIH-funded Tuberculosis Clinical Diagnostics Research Consortium (TB-CDRC) proposes to perform whole-genome sequencing of approximately 200 clinical *M. tuberculosis* strains from Uganda and South Korea that have a wide variety of drug-resistance patterns. The strains provide a unique resource because they have had their minimum inhibitory concentrations (MICs) to first (except pyrazinamide) and most second line drugs established in both liquid and solid media and have had breakpoint resistance re-confirmed by MGIT. The sequenced strains will also be deposited into BEI Resources for use by the scientific community. The goal is to facilitate development of new diagnostic and therapeutic approaches to tuberculosis (TB).

Solid culture based phenotypic testing remains the stalwart gold-standard for antimicrobial susceptibility testing. However, genotypic diagnostic tools have the potential to standardize and speed up the drug susceptibility testing process, while also reducing biohazard. Specifically, identification of the complement of mutations which do (and do not) encode for drug resistance will enable the community to develop diagnostic tests for detection of drug resistance that cover a broader spectrum of mutations and have better sensitivity and specificity compared to those currently available. In addition, the deposited, well characterized strain bank will provide a novel diagnostic testing panel for analyzing the competence of new tests and allowing for easy head-to-head comparison of all tests. This will speed up the test approval process and reduce the need for expensive early field-testing.

With respect to development of new anti-TB therapeutics, the production of an annotated SNP map linked with different drug resistance MICs will help establish novel resistance mechanisms and will facilitate studies to better understand the mechanisms of drug resistance evolution. This resource will enable the scientific community to develop novel therapeutic approaches to prevent the emergence of resistance in TB.
2. Justification

Provide a succinct justification for the sequencing or genotyping study by describing the significance of the problem and providing other relevant background information.

This section is a key evaluation criterion.

1. State the relevance to infectious disease for the organism(s) to be studied; for example the public health significance, model system etc.

The proposed project is highly relevant to the understanding of the biology of *M. tuberculosis* and to development of new diagnostic and therapeutic approaches.

With respect to TB diagnostics, the efficient and accurate detection of drug resistant tuberculosis (TB) is a major barrier to global TB control. Phenotypic methods for drug susceptibility testing are cumbersome, slow, and require extensive biosafety infrastructure. Molecular methods, based on detection of genetic mutations demonstrated to correlate with phenotypic resistance, are available for a limited set of mutations associated with resistance to some first- and second-line TB drugs. However, there are two critical issues that need to be addressed in order to realize the full potential of molecular-based drug susceptibility testing. The first issue is identification of the full spectrum of genetic changes associated with resistance to TB drugs. As an example, taken from the Table 1, the characterized *embB* mutation conferring ethambutol resistance is only associated with up to 65% of ethambutol resistant isolates. Generation of an annotated list of SNPs from a wide array of characterized resistance strains will help establish the remaining elusive causes of resistance. The second issue is the development and validation of clinical tests having mutation coverage sufficient to provide definitive, clinically actionable results. Table 2 provides examples of the current array of susceptibility tests available; suitable technologies are emerging, but progress towards the development of truly transformative TB resistance testing platforms has been hampered by the existing gaps in knowledge as to the full spectrum of mutations of interest, and absence of a readily accessible set of well-characterized *M. tuberculosis* strains for use in test development and validation. Availability of a set of genotypically and phenotypically well-characterized *M. tuberculosis* strains would be expected to stimulate and facilitate testing platform development, limit the need for early field trials, facilitate test approval processes, and provide a simple mechanism for direct comparison of new diagnostic kits.

With respect to novel therapeutic approaches to TB, the availability of annotated SNPs linked with different drug resistance minimum inhibitory concentrations (MICs) will facilitate studies to better understand the mechanisms of drug resistance evolution, which could result in new therapeutic interventions. The identification of new causes of drug resistance will also enable the scientific community to develop new therapeutics to counter drug-resistance development (for example pump inhibitors if pump regulation is determined to be associated with multi-drug resistance). Finally, the availability of this strain set will allow the scientific community to test new anti-TB drugs against a genetically well-characterized and diverse set of drug resistant mutants, which will speed the drug development process.
Table 1. Mechanisms of drug resistance

<table>
<thead>
<tr>
<th>Drug</th>
<th>Mechanism of action</th>
<th>Genes involved in resistance</th>
<th>Mode of resistance</th>
<th>Mutation frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isoniazid</td>
<td>Inhibition of mycolic biosynthesis and other effects</td>
<td><em>katG</em>, <em>inhA</em></td>
<td>Prevent prodrug conversion, Alter drug target</td>
<td>50-95, 8-43</td>
</tr>
<tr>
<td>Rifampin</td>
<td>Inhibition of RNA synthesis</td>
<td><em>rpoB</em></td>
<td>Alter drug target</td>
<td>95</td>
</tr>
<tr>
<td>Ethambutol</td>
<td>Inhibition of Arabinogalactan synthesis</td>
<td><em>embB</em></td>
<td>Alter drug target</td>
<td>47-65</td>
</tr>
<tr>
<td>Pyrazinamide</td>
<td>Depletion of membrane energy</td>
<td><em>pncA</em></td>
<td>Prevent prodrug conversion</td>
<td>72-97</td>
</tr>
<tr>
<td>Streptomycin</td>
<td>Inhibition of protein synthesis</td>
<td><em>rpsL</em>, <em>rrs</em>, <em>gidB</em></td>
<td>Alter drug target</td>
<td>52-59, 8-21, ?</td>
</tr>
<tr>
<td>Amikacin/kanamycin</td>
<td>Inhibition of protein synthesis</td>
<td><em>rrs</em>, <em>eis</em></td>
<td>Alter drug target</td>
<td>70-80, 22*</td>
</tr>
<tr>
<td>Capreomycin</td>
<td>Inhibition of protein synthesis</td>
<td><em>tlyA</em></td>
<td>Alter target modification</td>
<td>Rare</td>
</tr>
<tr>
<td>Quinolones</td>
<td>Inhibition of DNA gyrase</td>
<td><em>gyrA</em>, <em>gyrB</em></td>
<td>Alter drug target</td>
<td>75-94</td>
</tr>
<tr>
<td>Ethionamide</td>
<td>Inhibition of mycolic acid synthesis</td>
<td><em>etaA/ethA</em>, <em>inhA</em></td>
<td>Prevent prodrug conversion, Alter drug target</td>
<td>37, 56</td>
</tr>
<tr>
<td>Para-amino-salicylic acid</td>
<td>Inhibition of folic acid and iron metabolism?</td>
<td><em>thyA</em></td>
<td>Drug activation?</td>
<td>36</td>
</tr>
</tbody>
</table>

Table 2. Antimicrobial susceptibility testing methods
(based on Wilson, 2011, *Clin Infect Dis* 52, 1350-1355, and references therein)

<table>
<thead>
<tr>
<th>Antimicrobial susceptibility test</th>
<th>Drug resistance tested</th>
<th>Test method</th>
<th>Equipment</th>
<th>Sensitivity / Specificity (dataset)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agar proportion</td>
<td></td>
<td>Growth on solid media containing break-point drug concentration</td>
<td>BSL3</td>
<td></td>
</tr>
<tr>
<td>MDR-XDR TB color test</td>
<td>Isoniazid, rifampin, other</td>
<td>Growth on thin layer agar in quadrant petridish</td>
<td>BSL3</td>
<td>100% / 100% (small meta-analysis)</td>
</tr>
<tr>
<td>MODS (microscopic observation drug)</td>
<td>Isoniazid rifampin</td>
<td>Growth in liquid microtitre culture</td>
<td>BSL3, microscope</td>
<td>90% / 98.6%, 98% / 99.4% (pooled meta-analysis)</td>
</tr>
</tbody>
</table>
###susceptibility

<table>
<thead>
<tr>
<th>Method</th>
<th>Drugs</th>
<th>Reagents</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colorimetric assay e.g. Nitrate reductase assay</td>
<td>Isoniazid, rifampin, ethambutol, streptomycin, other</td>
<td>Color-change reagent added to liquid culture – detects metabolic activity of growing bacilli</td>
<td>BSL3, reagents 95.6% / 100%, 100% / 100%, 100% / 98.7%, 93.7% / 98% (small study)</td>
</tr>
<tr>
<td>TB-Biochip Microarray</td>
<td>Rifampin</td>
<td>Oligonucleotide microarray of RRDR of rpoB</td>
<td>BSL3, Microarray platform 80% / 100% (small study)</td>
</tr>
<tr>
<td>GenoType MTBDRplus</td>
<td>Rifampin, isoniazid</td>
<td>Line probe assay: PCR amplicons hybridized to probe strip, detect some rpoB, katG, inhA mutations</td>
<td>BSL3, PCR machine, reagents 99% / 99%, 96% / 100% (meta-analysis)</td>
</tr>
<tr>
<td>GenoType MTBDRsl</td>
<td>Fluoroquinolone, ethambutol, kanamycin, amikacin, capreomycin</td>
<td>Line probe assay</td>
<td>BSL3, PCR machine, reagents 87% / 96%, 57% / 92%, 77% / 100%, 100% / 100%, 80% / 98% (small study)</td>
</tr>
<tr>
<td>Xpert MTB/RIF</td>
<td>Rifampin</td>
<td>Real time PCR and molecular beacons detect rpoB mutations</td>
<td>Gene Xpert machine, reagents 99.1% / 100% (large field study)</td>
</tr>
</tbody>
</table>

2. **Are there genome data for organisms in the same phylum / class / family / genus?** What is the status of other sequencing / genotyping projects on the same organism including current and past projects of the NIAID GSC? Provide information on other characteristics (genome size, GC content, repetitive DNA, pre-existing arrays etc.) relevant to the proposed study. Have analyses been performed on the raw data already generated/published? If additional strains are proposed for a species, please provide a justification for additional strains?

The genome of *M. tuberculosis* has been sequenced and determined to be relatively large (4.4 Mbp), with a G+C content of 65.6% and 3,924 predicted coding sequences (Cole *et al.*, *Nature* **393**, 537-544). The *M. tuberculosis* genome is known to be non-plastic, such that any SNP mutations present in a resistant strain tend to have a corresponding selectable function (Saunders *et al.*, 2011, *J Infect* **62**, 212-217). Therefore it is expected that the generation of an annotated list of SNPs from a wide array of characterized resistance strains will help establish novel genotypic causes of drug resistance.

To the best of our knowledge from literature searches and the Broad database of sequenced genomes, at present there are no publicly available *M. tuberculosis* genome sequences of large strain sets with varied resistance phenotypes to both first and second line drugs with detailed MIC data attached. Conversely, there are no open-access collections of *M. tuberculosis* strains that have been genotypically characterized for drug resistance. We
propose that the CDRC *M. tuberculosis* strain collection is an optimal set for genotypic characterization as the complementary phenotypic data has already been comprehensively recorded and the collection will be deposited into BEI resources. The CDRC *M. tuberculosis* strain study collection is a large and intensively studied strain set that has been tested for first line (including pyrazinamide for most) and most second line phenotypic drug resistance by several methods including LJ proportions and MGIT, and eight-point MIC determinations for all first line anti-TB drugs (except pyrazinamide) and almost all second line drugs using TREK diagnostic Sensititre plates.

There are other proposed sequencing projects, such as the Broad’s Mycobacterium Comparative Database, that aim to sequence a small number of clinical strains with known clinical background, including susceptible/resistance determination, and there are also published genomic studies of select MDR/XDR strains (Ioerger, T. R. et al. 2009, *PLoS One* 4, e7778, Niemann et al., 2009, *PLoS One* 4, e7407). These studies will provide an interesting comparison, however, the scale of the CDRC TB project will provide a wealth of additional data about mutation frequency, low level/high level resistance and resistance evolution.

Various strains of *M. tuberculosis* have already been sequenced; these strains will be of great use in confirming the SNPs of the CDRC *M. tuberculosis* collection. The sequenced “wild-type” H37Rv strain will be used as the primary reference strain for establishing SNPs in the collection. However, once strain lineage is established, secondary confirmation of SNPs can be assessed with reference to sequenced susceptible strains of different *M. tuberculosis* lineages according to the Broad’s *M. tuberculosis* Diversity Database project. Finally, as a large body of sequence data will be generated, the SNP annotations can be compared to the collection itself for delineating corresponding MIC differences.

3. If analyses have been conducted, briefly describe utility of the new sequencing or genotyping information with an explanation of how the proposed study to generate additional data will advance diagnostics, therapeutics, epidemiology, vaccines, or basic knowledge such as species diversity, evolution, virulence, etc. of the proposed organism to be studied.

A limited subset of the strains in this set have been sequenced at widely known resistance loci for rifampin (*rpoB* core region), fluoroquinolones (*gyrA* resistance determining region) and *rrs* resistance encoding codons. A smaller set has been fully sequenced by the Global Novartis Foundation. However, most strains have not been sequenced or have had very limited sequencing performed. The proposed strategy would provide complete SNP identification using a uniform method on the complete strains study set. These mutations would allow the scientific community to confirm suspected associations between mutations and drug resistance, discover new mutations responsible for high or low-resistant MICs and identify the cause of drug resistance in subsets of strains that have wild type sequence at loci currently implicated in drug resistance. Additionally, this study would permit the identification of mutations specifically associated with multi-drug resistance (e.g. pump mutations) and mutations required to maintain virulence in drug-resistance strains. These last two goals addresses important research goals that have been largely inaccessible to the scientific community due to the requirement of large scale DNA sequencing.
3. Rationale for Strain Selection

4. Provide the rationale behind the selection of strains and the number of strains proposed in the study. The focus of the program is on potential agents of bioterrorism or organisms responsible for emerging or re-emerging infectious diseases. Non-select agents or non-pathogenic organisms will be considered when they can provide insight into these scientific areas.

We propose sequencing of approximately 200 clinical *M. tuberculosis* strains isolated from humans with TB disease. Approximately half of the strains originate from South Korea and half from Uganda. With respect to drug resistance profiles, the strains are heterogeneous, but within the set of 200 there are at least 50 strains resistant to each 1st and 2nd line anti-tuberculosis drug. As described above, this strain set that has been tested for first line (including pyrazinamide for most) and most second line phenotypic drug resistance by several methods including LJ proportions, Middlebrook 7H10 agar proportions, and MGIT, and eight-point MIC determinations for all first line anti-TB drugs (except pyrazinamide) and almost all second line drugs (streptomycin, amikacin, kanamycin, cycloserine, ofloxacin, moxifloxacin, PAS, ethionamide, rifabutin). These strains have been less well characterized genotypically.

With respect to *M. tuberculosis* phylogenetics, epidemiological data from the Republic of Korea have shown that the majority of strains are K family strains, derivatives of the Beijing family (Kang et al. *Journal of Medical Microbiology* 59, 1191-1197). In the Kampala, Uganda vicinity (where the Ugandan strains were collected), the T2 spoligotype family of strains has been shown to predominate. Therefore we anticipate that the phylogenetic distribution within the proposed strain set is reasonably restricted, which should facilitate sequencing data interpretation. Yet, the two strain sets (Ugandan versus Korean) are different from each other, suggesting that the data provided will be representative of the broader *M. tuberculosis* population.

4a. Approach to Data Production: Data Generation

5. State the data and resources planned to be generated. (e.g. draft genome sequences, finished sequence data, SNPs, DNA/protein arrays generation, clone generation etc.)

Data to be generated:
- A catalog of annotated SNPs for each strain
- A SNP based identification of strain phylogenetic lineage and sublineages
- A catalog of candidate SNPs associated with antibiotic resistance

Resources to be generated:
*M. tuberculosis* strains will be deposited at BEI resources for use by the scientific community.

4b. Approach to Data Production: Data Analysis
6. Briefly describe the analysis (value-add) envisioned to be performed subsequently by the community and the potential to develop hypotheses driven proposals given the datasets and resources produced by this work.

- Identifying the complement of mutations which do (and do not) encode for drug resistance will enable the community to develop diagnostic tests for detection of drug resistance that have better sensitivity and specificity compared to those currently available.

- Identifying new causes of drug resistance will enable the community to develop new therapeutics to counter drug-resistance development (for example pump inhibitors if pump regulation is found to be associated with multi-drug resistance).

- Producing annotated SNP with different drug resistance MICs will enable new studies to better understand the mechanisms of drug resistance evolution. This work could result in new therapeutic interventions.

- Making available a well characterized set of drug susceptible and drug resistant mutants will stimulate development activity and testing of new TB diagnostics, limit the need for early field trials, facilitate test approval processes, and provide a simple mechanism for direct comparison of new diagnostic kits.

- The availability of this strain set will also enable the community to test new anti-TB drugs against a genetically well-characterized and diverse set of drug resistant mutants. This will speed the drug development process.

5. Community Support and Collaborator Roles:

7. Provide evidence of the relevant scientific community’s size and depth of interest in the proposed sequencing or genotyping data for this organism or group of organisms. Please provide specific examples.

TB-CDRC investigators provide consultation to TB diagnostics developers, and in this capacity have met with lead scientists at Idaho Technologies, Akonni and parallel synthesis technologies who have identified the lack of a well-characterized MTB strain set as a barrier to their diagnostic development work. Further, the absence of information about genotype-phenotype correlations for pyrazinamide resistance is considered by the TB field generally to be a major barrier to selection of optimal treatment regimens for drug-resistant TB, since phenotypic testing for pyrazinamide resistance is inaccurate and technically complex.

8. List all project collaborators and their roles in the project

- Project lead: David Alland, M.D., Professor and Chair, Division of Infectious Diseases, University of Medicine and Dentistry of New Jersey; Scientific Director, TB-CDRC
Project investigator: Jerrold Ellner, M.D., Professor and Chair, Division of Infectious Diseases, Boston Medical Center; Principal Investigator, TB-CDRC

Project investigator: Susan Dorman M.D., Associate Professor of Medicine, Johns Hopkins University; co-Principal Investigator, TB-CDRC

Project investigator: Clifton Barry, Ph.D., Chief, Tuberculosis Research Section, NIAID, NIH

Project investigator: Moses Joloba, MbChB, Ph.D., Professor and Chair, Dept of Microbiology, Makerere University Medical School, Kampala, Uganda

Project coordinator: Susan Weir, Boston Medical Center

Karen Lacourciere, Ph.D., TB-CDRC Technical Project Officer, NIAID/DMID

9. List availability of other funding sources for the project.

The following will be supported by the NIAID/DMID TB–CDRC contract # HHSN272200900050C):

• Time and effort for personnel listed above
• Phenotypic drug susceptibility testing of *M. tuberculosis* strains
• Shipping of *M. tuberculosis* strains
• Preparation of DNA from *M. tuberculosis* strains

6. Availability & Information of Strains:

10. Indicate availability of relevant laboratory strains and clinical isolates. Are the strains/isolates of interest retrospectively collected, prepared and ready to ship?

Note: If samples are prospectively prepared the GSC can provide protocols and recommendation based on the Centers past experiences. The samples must however meet minimum quality standards as established by the Center for the optimal technology platform (sequencing/ genotyping) to be used in the study.

The approximately 200 *M. tuberculosis* strains are presently stored frozen at -80 °C and are ready to be shipped. DNA would be prepared at TB-CDRC clinical laboratories if appropriate from the Genome Sequencing Project perspective. Of note, each *M. tuberculosis* strain is derived from a bulk culture of a clinical specimen; strains are not derived from single colonies and therefore we cannot exclude the possibility that some of the strains are mixtures. However, since the phenotypic drug susceptibility testing was performed on the existing stored material, we believe that the most appropriate scientific approach is to perform sequencing on the phenotypically-characterized material, and use sequence analysis techniques to assess for heterogeneity within each sample.
11. What supporting metadata and clinical data have been collected or are planned on being collected that could be made available for community use?

As above, drug susceptibility testing using the conventional Middlebrook agar proportions method and minimum inhibitory concentration results using a liquid testing kit (TREK MYCOTB MIC plate) are underway for isoniazid, rifampin, rifabutin, amikacin, streptomycin, cycloserine, ethambutol, ethionamide, kanamycin, moxifloxacin, ofloxacin, and PAS; and previously the strains have been tested using MGIT and LJ DST. These results will be made available to the community.
7. Compliance Requirements:
7a. Review NIAID’s Reagent, Data & Software Release Policy:
NIAID supports rapid data and reagent release to the scientific community for all sequencing and genotyping projects funded by NIAID GSC. It is expected that projects will adhere to the data and reagent release policy described in the following web sites.
http://www3.niaid.nih.gov/LabsAndResources/resources/mscs/data.htm
Once a white paper project is approved, NIAID GSC will develop with the collaborators a detailed data and reagent release plan to be reviewed and approved by NIAID.

Accept ☒ Decline □

7b. Public Access to Reagents, Data, Software and Other Materials:

12. State plans for deposit of starting materials as well as resulting reagents, resources, and datasets in NIAID approved repositories. Sequencing projects will not begin until the strain is deposited into NIAID funded BEI repository (http://www.beiresources.org/). This includes web based forms are completed by the collaborator and received by the NIAID BEI (http://www.beiresources.org/).

The M. tuberculosis strains will be deposited with BEI. Datasets will be deposited in the most appropriate repository as determined by NIAID.

7c. Research Compliance Requirements

Upon project approval, NIAID review of relevant IRB/IACUC documentation is required prior to commencement of work. Please contact the GSC Principal Investigator(s) to ensure necessary documentation are filed for / made available for timely start of the project.

Investigator Signature:

[Signature]

Investigator Name: David Alland, M.D. Date: 24 May 2012