Plan Overview

A Data Management Plan created using DMPTool

DMP ID: https://doi.org/10.48321/D14D0B

Title: Advancing genomic, transcriptomic and functional approaches to combat globally important and emerging pathogens

Creator: Daniel Park - **ORCID:** 0000-0001-7226-7781

Affiliation: Broad Institute (broadinstitute.org)

Principal Investigator: Bruce Birren, Daniel Neafsey

Data Manager: Daniel Park

Project Administrator: Rachel Newmiller

Funder: National Institute of Allergy and Infectious Diseases (niaid.nih.gov)

Funding opportunity number: RFA-AI-23-015

Template: NIH-Default DMSP

Project abstract:

Infectious diseases remain an important global threat despite the continuing invention of drugs, vaccines, and other effective interventions. Two principal factors drive this continued threat: 1) efficacious interventions are nearly always met with evolutionary responses on the part of pathogens and disease-transmitting vectors, reducing their potency over time, and 2) the threat of infectious diseases is dynamic and exacerbated by the growing and increasingly connected human population, as evidenced by disruptive outbreaks, epidemics and pandemics in recent years driven by Zika, Ebola, *Candida auris*, antimicrobial-resistant bacteria, and of course SARS-CoV-2.

Genomics, transcriptomics, and associated scientific tools for investigating pathogens in precise detail and at tremendous scale are a key capacity for maintaining control over existing and emerging infectious diseases. Our GCID will continue to make use of a highly experienced management team leading our Administrative and Technology and Data Cores, and a proven team of scientists well versed in genomic analysis and microbial research, to apply innovative genomic approaches to solve crucial problems in infectious disease research. Our four Research Projects, devoted to viral, bacterial, fungal, and parasitic diseases, and vectors responsible for their transmission, will target high priority pathogens and pathogen-host-vector systems with a major impact on the global burden of disease. To maximize synergy between Projects, our research aims were developed with a

common conceptual framework and shared approaches, and leverage working relationships that have been developed through years of close collaboration and scientific dialogue. Using those shared concepts and approaches, we will pursue **three broad, long-range objectives** in our GCID across our four Projects:

Specific Aim 1: To identify and profile novel threats and therapeutic targets. We hypothesize that informative genomic methods to gather and prioritize information will be required to counter the emergence of resistance to existing therapeutics and respond effectively to new disease emergencies and spillover events.

Specific Aim 2: To inform clinical treatments for the benefit of patients and populations. We hypothesize that genomic and transcriptomic data can improve disease treatment by establishing the basis of resistance and virulence.

Specific Aim 3: To drive precision public health surveillance and responses to disease emergence. We hypothesize that genomic data can productively inform awareness and responses to infectious threats to support public health.

Impact: We will develop innovative genomic approaches and datasets to address critical gaps in knowledge and capacity across diverse infectious diseases, and disseminate data and tools to the research community.

Start date: 04-01-2024

End date: 04-01-2029

Last modified: 01-18-2024

Copyright information:

Advancing genomic, transcriptomic and functional approaches to combat globally important and emerging pathogens

This proposal will produce large scale genomic data obtained from pathogenic microbes that were sampled from human patients, laboratory isolates and cultures, animal and insect hosts, and environmental sampling. The data generated will include microbial genomic sequence (DNA and RNA), metagenomic sequence, and transcriptomic / expression data. These may be generated on bulk samples, single-cell, or some other spatially separated sequencing approach. These may include targeted/enriched sequences or agnostic sequences.

In addition to raw sequence data, the Center will produce associated analysis outputs, including assembled genomes, genotype/strain calls and expression profiles.

Raw reads: microbial DNA and RNA sequence reads generated under this proposal will be shared. These data will also include information on sequencing templates and quality values as well as primer sets (if applicable) and laboratory methods. Raw sequence data will first be informatically filtered for any human-mapping reads prior to sharing, so as to minimize identifiability risk to patients.

Assemblies and annotations: assembled microbial genomes and predicted gene annotations will be shared after Center personnel have completed consistency checks and quality control.

Gene expression data: RNA-Seq and other transcriptomic or gene expression data sets generated by the Center will be shared after passing quality checks after being derived from primary data.

Relevant metadata (clinical data or other sample-associated data) that are essential for the biological interpretation of genome sequence data will be shared. These often will include sample collection dates, locations, specimen types and collection mechanisms.

For clinical specimens, clinical metadata such as disease, symptoms, or outcomes, as permitted by each study may be included and shared.

If in vitro phenotypic information is generated by the Center on microbial isolates (e.g. drug susceptibility, MICs, etc), these will also be shared as metadata.

Genomic data types (short read, assemblies, etc) shared in standard public repositories at NCBI have a well established ecosystem of freely available bioinformatic tools that can be used to access or manipulate the data. Such data is downloadable from NCBI in fastq, bam, fasta, or vcf formats, all of which are open formats and do not require specialized tools to access. There are no plans to share any data via proprietary data formats.

To the extent that our Center develops novel analytic methods for more effective analysis, assembly, or interpretation of such data, those methods would be distributed in an open source manner by the Center's Technology and Data Core via the mechanisms described in its Specific Aims and Research Strategy.

Data will be stored and shared in common and open formats.

For raw sequence data, this may include fastq or bam formats, however, the APIs and tools for accessing the NCBI SRA database (where we intend to share it) allow data consumers to specify other standard file formats of their choice and perform on-the-fly file format conversion while downloading.

Genomes may be distributed in the open fasta format. Gene annotations may be distributed in GFF or NCBI's TBL formats--NCBI's APIs allow data consumers to select other standard file formats of their choice for gene

annotations.

Sample metadata will conform to the templates and requirements provided by the NCBI BioSample database. These will frequently utilize the NCBI *Microbe* (bacterial), *Pathogen.cl* (clinical) or *Pathogen.env* (environmental) metadata templates, however there are other project-specific templates provided by BioSample that might be used. For metadata fields where NCBI BioSample does not enforce a controlled vocabulary, we will, where possible, utilize community standard ontologies and vocabularies relevant to the pathogen being studied, such as those defined by the Public Health Alliance for Genomic Epidemiology (PHA4GE) or the Genomic Standards Consortium Minimum Information about any Sequence (MIxS) templates.

Datasets will primarily be shared via NCBI's databases, which are maintained by the NIH and globally replicated by the INSDC. This will maximize discoverability, accessibility, and longevity of the data.

Raw reads: raw DNA and RNA sequence reads generated under this proposal will be submitted to the Short Read Archive (SRA) at NCBI/NLM/NIH after being filtered for human-mapping reads. These data will also include information on sequencing templates and quality values as well as primer sets (if applicable) and laboratory methods, utilizing SRA's standard metadata model.

Assemblies and annotations: assembled microbial genomes and associated gene annotations will be made available via NCBI's Nucleotide (a.k.a. Genbank) and/or Assembly databases. The appropriate database will be selected based on community standards for that pathogen.

Gene expression data: RNA-Seq, single cell, and other transcriptomic or gene expression data sets generated by the Center will be submitted to the public database at NCBI's Gene Expression Omnibus (GEO).

Relevant metadata that are essential for the biological interpretation of sequence data will be made available to the scientific community through NCBI's BioSample database. The above data elements (SRA, Genbank, GEO) will be linked with their associated BioSample records.

For certain pathogens or data sets, the above data may additionally be linked or replicated in pathogen-specific portals, such as those provided by the NIAID Bioinformatics Resource Centers.

All primary data will be indexed in NCBI's databases as described above. These repositories are supported and maintained by the NIH/NLM/NCBI and globally replicated by the INSDC in Europe/UK and Japan. All individual data elements receive accession numbers in each database -- these are then linked across databases (e.g. SRA, GEO, Genbank) via a BioSample accession, which is uniquely searchable in any INSDC database. All data elements associated with a particular data set will be connected with a unique BioProject ID. All data sets released by this Center will have their BioProjects linked to this NIH grant number: U19AI110818.

All genomic data will be publicly shared as rapidly as possible via NCBI databases, but at minimum, within the timeframes specified by NIAID in RFA-AI-23-015 which are listed below for each data type. No embargo periods will be imposed prior to public data release that would exceed the timeframes below. For Emergency Response Projects, all data is expected to be publicly released much sooner than the guaranteed timeframes below. All published data is expected to remain available for the full lifespan of NCBI & INSDC.

Raw reads: raw DNA and RNA sequence reads generated under this proposal will be submitted for public release as rapidly as possible within 45 calendar days of the completion of passing quality control checks after data generation.

Assemblies and annotations: assembled microbial genomes and predicted gene annotations will be submitted for

public release within 45 calendar days of assembly and annotation generation and validation, assuming no significant validation or quality control errors.

Gene expression data: RNA-Seq and other transcriptomic or gene expression data sets generated by the Center will be publicly released within nine months of passing quality checks after being derived from primary data.

All shared metadata will be submitted for public release concurrently with, and linked to, the relevant primary data type (raw reads, assemblies, expression data).

There are no anticipated factors or limitations that will affect the access, distribution or reuse of the scientific data generated and shared by the proposal.

This project does not expect the need to release data in controlled-access repositories. Microbial data that is shared will be shared by unrestricted download from public access repositories. If required by NIAID as part of our Collaborative Agreement, or if appropriate as part of an Emergency Response Project, potentially identifying human data or otherwise sensitive genomic data or metadata may be deposited in a controlled-access database as designated by NIAID. However, this is not anticipated at the time of this proposal.

Relevant metadata (clinical data or any other type of data) that are essential for the biological interpretation of genome sequence data, will be publicly released in a de-identified manner in association with microbial genomic data. However, any metadata that may potentially identify human subjects will not be released in openly accessible public databases unless permitted by the governing IRB or the Broad Office of Research Subjects Protection. Such metadata may be released at degraded resolution if necessary.

The Technology and Data Core co-lead, Daniel Park, ORCID: 0000-0001-7226-7781, will be responsible for the day-to-day oversight of lab/team data management activities and data sharing. Broader issues of DMS Plan compliance oversight and reporting will be handled by the Center PIs, Project/Core Leads, and Admin Core, as part of general stewardship, reporting, and compliance processes.