Plan Overview

A Data Management Plan created using DMP Tool

DMP ID: https://doi.org/10.48321/D12946

Title: Virome Investigation in Diverse Human Populations

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

Affiliation: Broad Institute (broadinstitute.org)

Principal Investigator: Pardis Sabeti, Dong Wang

Data Manager: Daniel Park

Project Administrator: Elizabeth Curtis

Funder: National Institutes of Health (nih.gov)

Funding opportunity number: RFA-RM-23-019

Grant: U54AG089325

Template: NIH-Default DMSP

Project abstract:

The human virome – the collection of all the viruses that are found inside and on the human body across multiple anatomical sites – plays a significant role in health, with many viruses being non-harmful or even beneficial. However, the composition, dynamics, and variability of these viruses remains poorly understood. . The goal of our Human Virome Program (HVP) Variant Characterization Center (VCC) is to uncover and curate the spectrum of viruses present in humans and characterize key features of their biology and dynamics, including mapping viral tropism, monitoring changes over time, identifying differences in groups of viruses, and cataloging all

phages. The VCC will leverage a combination of genomics, transcriptomics, imaging, and computational techniques to develop a systematic, multimodal approach for understanding the human virome in precise detail and at scale. The project will leverage a specimen collection from five large, longitudinal and diverse cohorts that share a common study design, collectively encompassing >350,000 participants across all 55 U.S. states and territories. From it, our Biospecimen Collection Core (BCC) will select a diverse 4,000-participant cohort of varying ages, sex/gender, geographic origin, and racial/ethnic diversity, which will be analyzed by the Biospecimen Analysis Core (BAC) using existing cutting-edge multi-omics approaches, and where needed developing and adapting new methods to improve viral genome recovery and characterization. The Data Analysis and Submission Core (DASC) will use state-of-the-art computational approaches to process, store, and analyze these data. The Ethical Legal and Societal Implications (ELSI) core will manage the responsible engagement with participants and explore how results and findings can be ethically communicated to participants. The VCC, managed by the Broad Institute of MIT and Harvard, Brigham and Women's Hospital, and Harvard T.H. Chan School of Public Health, emphasizes effective leadership and collaboration among its five cores. It will be led by a team of experts in diverse areas, including viral genomics, the human microbiome, bioinformatics, technology development, and management of large cohort studies. An Administrative Core (AC) will ensure clear milestones, defined metrics, and contingency plans to achieve its goals. The timeline encompasses sample retrieval, participant engagement, and data processing over multiple years, with ongoing evaluation and proactive measures to identify and address potential challenges. The AC will also implement a Plan to Enhance Diverse Perspectives (PEDP) for the study team and participants. The VCC will be committed to transparently sharing resources, including data, reagents, protocols, and tools, through public repositories, journals, and other platforms. Through the project, we will develop an innovative, multimodal approach to characterizing the human virome, fill gaps in understanding of how the virome influences human health, and disseminate tools and data broadly to the scientific community.

Start date: 07-01-2024

End date: 06-30-2029

Last modified: 02-19-2025

Copyright information:

Virome Investigation in Diverse Human Populations

Data Type

Types and amount of scientific data expected to be generated in the project: *Summarize the types and estimated amount of scientific data expected to be generated in the project.*

Describe data in general terms that address the type and amount/size of scientific data expected to be collected and used in the project (e.g., 256-channel EEG data and fMRI images from ~50 research participants). Descriptions may indicate the data modality (e.g., imaging, genomic, mobile, survey), level of aggregation (e.g., individual, aggregated, summarized), and/or the degree of data processing that has occurred (i.e., how raw or processed the data will be)

This proposal will produce large scale genomic, transcriptomic, and imaging data obtained from microbes that were sampled from human participants. 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.

In addition, phenotypic data associated with biospecimen collections will be generated by this proposed project.

Scientific data that will be preserved and shared, and the rationale for doing so: *Describe which scientific data from the project will be preserved and shared and provide the rationale for this decision.*

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.

Metadata, other relevant data, and associated documentation: Briefly list the metadata, other relevant data, and any associated documentation (e.g., study protocols and data collection instruments) that will be made accessible to facilitate interpretation of the scientific 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 health conditions and medication use, as permitted by each study may be included and shared. This also includes sample identifiers, participant age, sex, anthropometrics, demographic information, socioeconomic status, health conditions, family histories of diseases, lifestyles (e.g., diet, smoking, and physical activity), and sociopsychological factors. Study protocols and questionnaires of the participating studies either have been published or will be published.

I*n vitro* phenotypic information is generated by the Center on specimens (e.g. plasma biomarkers, drug susceptibility, MICs, etc), these will also be shared as metadata.

Related Tools, Software and/or Code

State whether specialized tools, software, and/or code are needed to access or manipulate shared scientific data, and if so, provide the name(s) of the needed tool(s) and software and specify how they can be accessed.

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 Data Analysis and Submission Core via the mechanisms described in its Research Strategy.

Standards

State what common data standards will be applied to the scientific data and associated metadata to enable interoperability of datasets and resources, and provide the name(s) of the data standards that will be applied and describe how these data standards will be applied to the scientific data generated by the research proposed in this project. If applicable, indicate that no consensus standards exist

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 similar

standardized 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.

Data Preservation, Access, and Associated Timelines

Repository where scientific data and metadata will be archived: Provide the name of the repository(ies) where scientific data and metadata arising from the project will be archived; see Selecting a Data Repository)

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 viral species.

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.

Phenotypic data: protected patient data including protected phenotypic and/or demographic data will be stored at NCBI's dbGaP database.

The Center's Data Analysis and Submission Core will additionally work closely with the HVP CODCC to develop and contribute data towards HVP-specific portals, data repositories, and similar resources.

How scientific data will be findable and identifiable: Describe how the scientific data will be findable and identifiable, i.e., via a persistent unique identifier or other standard indexing tools.

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 our assigned NIH grant number.

When and how long the scientific data will be made available: Describe when the scientific data will be made available to other users (i.e., no later than time of an associated publication or end of the performance period, whichever comes first) and for how long data will be available.

All genomic data will be publicly shared on NCBI within the timeframes specified below for each data type. No embargo periods will be imposed prior to public data release that would exceed the 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).

Access, Distribution, or Reuse Considerations

Factors affecting subsequent access, distribution, or reuse of scientific data: NIH expects that in drafting Plans, researchers maximize the appropriate sharing of scientific data. Describe and justify any applicable factors or data use limitations affecting subsequent access, distribution, or reuse of scientific data related to informed consent, privacy and confidentiality protections, and any other considerations that may limit the extent of data sharing. See Frequently Asked Questions for examples of justifiable reasons for limiting sharing of 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.

Whether access to scientific data will be controlled: State whether access to the scientific data will be controlled (i.e., made available by a data repository only after approval).

This proposal does not anticipate the need to utilize controlled-access databases for large scale 'omics data--all of these data types will utilize NCBI/INSDC public databases.

However, certain individual-level metadata (e.g. demographic, phenotypic) may, due to privacy and identifiability concerns, be de-precisioned or entirely redacted from public release in NCBI BioSample. Such identifiable metadata may, instead, be released via a controlled access ecosystem managed by a Data Access Committee mechanism. This may include NCBI’s dbGaP or similar mechanisms. We will

work closely with the HVP CODCC to standardize the mechanism for controlled access across the HVP.

Protections for privacy, rights, and confidentiality of human research participants: If generating scientific data derived from humans, describe how the privacy, rights, and confidentiality of human research participants will be protected (e.g., through de-identification, Certificates of Confidentiality, and other protective measures).

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 may be withheld from open public release or released at degraded resolution if appropriate.

Specifically: in order to minimize risks to study participants, data submitted to the NIH data repository will be de-identified and coded using a random, unique code. Data will be de-identified according to the following criteria: 1) the identities of subjects cannot be readily ascertained or otherwise associated with the data by the repository staff or secondary data users (45 C.F.R. 46.102(f)); 2) the identifiers enumerated in section 45 C.F.R. 164.514 (b)(2), the HIPAA Privacy Rule, are removed; and 3) the submitting institution has no actual knowledge that the remaining information could be used alone or in combination with other information to identify the subject of the data. Keys to codes will be held by the submitting institutions. Submissions of sequence data will be accompanied by a written certification. All submissions to the NIH data repository will be accompanied by a certification by the responsible Institutional Official(s) of the submitting institutions that they approve submission to the NIH data repositories. The certification will assure that: 1) the data submission is consistent with all applicable laws and regulations, as well as institutional policies; 2) the appropriate research uses of the data and the uses that are specifically excluded by the informed consent documents are delineated; 3) the identities of research participants will not be disclosed to the NIH data repositories; 4) an IRB and/or Privacy Board, as applicable, reviewed and verified that the submission of data to the NIH data repository and subsequent sharing for research purposes are consistent with the informed consent of study participants from whom the data and samples were obtained; the investigator’s plan for de-identifying datasets is consistent with the standards outlined above; it has considered the risks to individuals, their families, and groups or populations associated with data submitted to the NIH data repository; and, the sequence and phenotype data to be submitted were collected in a manner consistent with 45 C.F.R. Part 46.

Oversight of Data Management and Sharing

Describe how compliance with this Plan will be monitored and managed, frequency of oversight, and by whom at your institution (e.g., titles, roles).

The Data Analysis and Submission Core 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, Core Leads, and Admin Core, as part of general stewardship, reporting, and compliance processes.

The Research Data Protection Office and the Research Management Office at Brigham and Women’s Hospital, responsible for IRB oversight of this award, have developed a data management and sharing plan compliance system as part of their process for submitting the annual NIH progress report. The offices will be monitoring the submission of data to the NIH data repositories.