Plan Overview

A Data Management Plan created using DMP Tool

DMP ID: https://doi.org/10.48321/D16W8T

Title: Characterizing genomic variation in Neanderthals and Denisovans and its functional impact on modern human populations

Creator: Fernando Villanea guevara - **ORCID:** 0000-0002-6661-0368

Affiliation: University of Colorado Boulder (CU Boulder) (colorado.edu)

Principal Investigator: Fernando Villanea Guevara

Data Manager: Fernando Villanea Guevara

Project Administrator: Fernando Villanea Guevara

Funder: National Institutes of Health (nih.gov)

Funding opportunity number: PAR-23-145

Grant: https://grants.nih.gov/grants/guide/pa-files/PAR-23-145.html

Template: NIH-Default DMSP

Project abstract:

There are now hundreds of ancient genomes available from a wide range of species, including extinct archaic humans: Neanderthals and Denisovans. Such data provide a direct window into the history of demography and natural selection in the recent past, and have greatly contributed to the understanding of how humans adapted to new environments after expanding outside of Africa. Direct comparisons of archaic and modern human genomes have revealed a complex landscape of genetic admixture, where living humans today carry a small but significant proportion of archaic

DNA. This archaic inheritance affects the fitness and health of living people. While bioinformatics methods for identifying archaic genome variants are very advanced, functional validation of how archaic genome variants differ from modern human equivalents in vitro lags behind. The proposed research will close this gap by performing functional validation experiments leveraging state-of-the-art cellular assays that take advantage of high-throughput DNA sequencing. Moreover, the focus of these experiments is in understanding the evolutionary history and medical consequences of archaic ancestry in underrepresented populations: in particular, Indigenous American and Latino individuals, whose genome ancestry is complex following 500 years of European colonization. We will focus on pharmacogenes, genes responsible for the metabolizing of exotic substances, which have a direct link to both adaptation to novel environments, and metabolizing of plant-derived medical drugs in modern medicine. By working with collaborators who specialize in the health consequences of pharmacogene variation in Indigenous populations, we will produce and disseminate knowledge in an ethical, responsible, and inclusive manner. We will also develop methodology to elucidate details of how modern humans and archaic humans, such as Neandertals interbred. By leveraging patterns of variation in Neandertal DNA sharing within and among European, East Asian, and Southeast Asian populations, we will determine the number of times that humans and Neandertals interbred, with the particular research question did Neanderthals interbred with modern humans in the Indian subcontinent.

Start date: 07-01-2024

End date: 07-01-2029

Last modified: 07-08-2024

Copyright information:

Characterizing genomic variation in Neanderthals and Denisovans and its functional impact on modern 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 project will produce genotype calls for Neanderthal ancestry for the 1000 Genomes Project genomes. The genotype calls will be generated through a computational pipeline. Genotype calls will be produced for up to 2504 haploid genomes (the total number of haploid genomes in the 1000 Genomes Project). The following data files will be produced in the course of the project: BED files containing positions of the genome detected to be of Neanderthal origin.

This project will produce expression data and variant data for CYP450 genes. The data will be generated through sequencing of single cells and processed using a computational pipeline. This data will be collected from a minimum of 11 independent experiments, with each independent experiment representing a separate CYP450 gene. The following data files will be produced in the course of the project: TSV files containing variant data, and TSV files containing positional data.

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.*

All data produced in the course of the project will be preserved and shared.

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.

To facilitate interpretation of the data, computational scripts containing the complete analyses pipelines will be created, shared, and associated with the relevant datasets.

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.

Bespoke computer scripts in Python will be created for data analyses. All scripts will be hosted and made freely available in the Villanea Lab GitHub account.

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

Whenever possible, we will use standard data formats such as BED or VCF files to structure and organize our data.

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.

All dataset(s) that can be shared will be deposited in the Villanea Lab Github along with instructions and computer scripts needed to replicate all results.

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 associated GitHub repositories will include README files detailing metadata.

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.

Shared data generated from this project will be made available as soon as possible, and no later than the time of publication or the end of the funding period, whichever comes first. The duration of preservation and sharing of the data will be permanent.

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.

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

Controlled access will not be used. The data that is shared will be shared by unrestricted download

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).

Question not answered.

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).

Lead PI Dr. Fernando Villanea, ORCID: 0000-0002-6661-0368, 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 PI as part of general University of Colorado Boulder stewardship, reporting, and compliance processes.
