s

# DATA MANAGEMENT AND SHARING PLAN

If any of the proposed research in the application involves the generation of scientific data, this application is subject to the NIH Policy for Data Management and Sharing and requires submission of a Data Management and Sharing Plan. If the proposed research in the application will generate large-scale genomic data, the Genomic Data Sharing Policy also applies and should be addressed in this Plan. Refer to the detailed instructions in the application guide for developing this plan as well as to additional guidance on [sharing.nih.gov.](https://sharing.nih.gov/) The Plan is recommended not to exceed two pages.

Text in italics should be deleted. There is no “form page” for the Data Management and Sharing Plan. The DMS Plan may be provided in the *format*

shown below.

An example DMS plan proposing to collect genomic, phenotypic, and clinical data from human subjects.

# Element 1: Data Type

1. **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.

Our genomic study will generate whole genome sequencing data, phenotypic and clinical data for 500 research subjects, which will be deidentified and deposited in an NIH-designated data repository using the standard formats (FASTQ, TSV formats) as well as the associated metadata (TXT format). The expected size of the data to be generated is about 25 TB.

# 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 raw and processed genomics files and all clinical and phenotypic data will be shared. Even though, we share the processed data (such as BAM files, VCF files, etc.), it is important to share the original raw data with the community to enable analysis using newer tools that would be available in the future.

# 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.

The Institutional Certification will be submitted to NIH during the dbGaP registration process once we have been told that a grant award is likely. Within the first six months following the award, we will submit the Data Submission Agreement to {Name of the NIH Institute funding this project}. A brief study protocol with corresponding metadata fields will also be submitted at that time and will be made freely available.

# Element 2: 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.

Genotypic data undergo an extensive automated data cleaning process in the laboratory. Our replication plan for observed associations is outlined in the Research Strategy. While all sequencing data from this proposal will be generated using {Illumina} instrumentation, differences in read depth and primer libraries between studies will require joint re-calling of all genotypes from raw read files to yield the highest possible quality calls and a harmonized dataset for future use in follow-up and unrelated studies. Using the Broad Institute’s Genome Analysis Toolkit (GATK), we will apply standard Best Practices workflows for single nucleotide variant (SNV) and Indel discovery from whole genome sequence alignment files (SAM/BAM). These steps should ensure that final association results are representative of “true” genotypes rather than miscalls or confounded genotypes that are unlikely to replicate in independent populations.

# Element 3: 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.

We will adhere to the community standards developed by the Genomic Data Commons (GDC/NCI/NIH) for data types and file formats for clinical, biospecimen, and molecular data. GDC data standards on data elements, data dictionaries, data types and file formats are accessible on the GDC website. Data will be prepared and submitted according to GDC-specific XML, TSV, or JSON formats, and molecular sequencing data will be submitted using the industry standard data formats such as BAM and FASTQ, and variant calling data in VCF format.

The following common data elements will be collected to facilitate aggregation of this data set with other data sets. At the sample level, we will provide Age, Sex at Birth and the Source Tissue. At the experiment level, the description of the genomic experiment will be provided using the variable such as Library\_Strategy (whole genome or whole exome sequencing), Library\_Source (genomic or metagenomic), Library\_Kit (Agilent Sureselect XT or Clinical Research Exome Kit), Library\_Selection (PCR), Library\_Layout (single- or paired-end sequencing), Sequencing\_Platform (Illumina or other) and Instrument\_Model (Hiseq, MySeq or Novoseq). As described in the Research Plan, the additional phenotypic and clinical information will be collected using the following data dictionaries.

# Element 4: Data Preservation, Access, and Associated Timelines

1. **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](https://sharing.nih.gov/data-management-and-sharing-policy/sharing-scientific-data/selecting-a-data-repository)).

All data will be deposited to SRA starting 12 months after the award begins and will be deposited every six months thereafter following the usual data submission dates.

# 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.

Our genomic study will be [registered with dbGaP](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?document_name=HowToSubmit.pdf), and our raw whole genome sequencing data and derived data will be submitted to Sequence Read Archive (SRA). Data will be findable for the research community through the NIH supported BioProject portal using the BioProject ID as a persistent unique identifier. From the BioProject page, links to access sample level data from BioSample and SRA repositories (both are supported by NIH/NLM) will be available. In addition, the dbGaP study also points to these data to help researchers find the data using the BioProject ID. At the time of publication, the BioProject ID will be provided to the publisher, which is a requirement for most of the journals these days. Also, links to the BioProject ID will be provided in the supplementary data files of the publications.

# 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.

The research community will have access to data at the end of the grant award or when a publication has been submitted, whichever comes first. Once the data are submitted to SRA and linked to dbGAP, it will be accessible the research community permanently. We will comply with the broad and rapid data sharing as outlined in the NIH Genomic Data Sharing Policy and will adhere to the future changes to NIH-wide data sharing policies.

**Element 5: Access, Distribution, or Reuse Considerations**

1. **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](https://sharing.nih.gov/faqs%23/data-management-and-sharing-policy.htm) [Asked Questions](https://sharing.nih.gov/faqs%23/data-management-and-sharing-policy.htm) for examples of justifiable reasons for limiting sharing of data.

All research participants will be consented for broad data sharing.

# 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).

All data access will be permitted as per the NIH Genomic Data Sharing Policy guidelines. To request access of the data, researchers will use the standard processes outlined by the dbGAP.

# 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).

An institutional IRB approval will be obtained before collecting data from human subjects. Consent will be obtained from all research participants for research use of genomic data either in a de-identified or identified format, as relevant to the project. Samples consented only for the de-identified usage will be processed differently from those consented for identifiable use and stored on a separate server. All the HIPAA variables will be removed, and de-identified patient IDs will be created using a one-way hash system. Only the de-identified patient IDs will be used in all data files containing phenotypic or clinical data, which will be used for sharing with the research community. All data will be stored on secure servers located at the HIPAA-compliant data center at the University of Nebraska Medical Center campus that maintains strict enterprise-level firewall security measures.

# Element 6: 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 Office of Sponsored Programs at University of Nebraska Medical Center administering this award has created a data management and sharing plan compliance system as part of their process for submitting this application. In addition, the following individuals will monitor and manage the implementation of this Plan on a day-to-day basis:

Lead PI, Jane Doe PhD, ORCID: xxxx-xxxx-xxxx-xxxx, will be responsible for the day-to-day oversight of data management activities and data sharing. Broader issues of DMS Plan compliance oversight and reporting will be handled by the PI and Co-I team as part of general stewardship, reporting, and compliance processes. The following individuals will be responsible for data collection, management, storage, retention, and dissemination of project data, including updating and revising the Data Management and Sharing Plan when necessary

John Doe, Database manager, UNMC, ORCID 0000-000x-xxxx-xxxx, johndoe@unmc.edu, will be responsible for…