***GRC Whole Genome/Whole Exome DMP template***

**Element 1: 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 whole genome sequencing data from <insert experiment-specific info regarding species and conditions>. Data will be collected from <insert number of samples> specimens, generating datasets totaling approximately <insert expected total data size: estimate ### per sample for WGS and ### per sample WES> Gb in size. The data files produced in the course of the project includes: fastq sequence files, unmapped bam files, cram files, and variant call files <If human data: mention steps to de-indentify participant 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.**

The following data produced in the course of the project will be preserved and shared: Raw sequencing files and variant call files. These files will allow other researchers to use the data and reproduce the analysis.

**A brief listing of 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, experimental metadata describing the experimental condition for each sample as well as relevant information associated with the samples will be included. The metadata will be included in tab separated file (tsv). <Insert list of experiment-specific metadata categories here such as the experimental conditions, batches, and/or relationships between the samples>.

**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**

Raw sequencing files in fastq format will be made available, these raw files can be converted to unmapped bam files (uBAM files) using Picard, and then can be used as input to gatk4 pipeline for variant detection. Fastq files do not require any specialized tools to access. Variant call files do not require any special tools to access, and can be manipulated, annotated and/or filtered using tools such as open cravat or on the Linux command line. Metadata detailing relevant sample information will be made available in tsv format and will not require the use of specialized tools to be accessed.

**If applicable, specify how needed tools can be accessed, (e.g., open source and freely available, generally available for a fee in the marketplace, available only from the research team) and, if known, whether such tools are likely to remain available for as long as the scientific data remain available.**

The following tools are used in this analysis along with the current version we plan to use. All of the following tools are freely available:

|  |  |  |
| --- | --- | --- |
| **Tool**  | **Version**  | **URL**  |
| Picard | 2.22.1 |  https://github.com/broadinstitute/picard/releases/tag/2.27.5 |
| Gatk4 |  4.1.8.0 |  https://github.com/broadinstitute/gatk/releases |
| Open Cravat |  2.2.7 | https://github.com/KarchinLab/open-cravat-readthedocs/blob/master/quickstart-command-line.rst |

**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**

The FAIR Data sharing protocols will be applied, so that the data will be Findable, Accessible, Interoperable, and Re-usable. Our sequencing data will be structured and described using the following description of standards: All shared studies contain 1) The description of the biological system, samples, and the experimental variables being studied, 2) The sequence read data for each assay, 3) The ‘final’ processed (or summary) data for the set of assays in the study, 4) General information about the experiment and sample-data relationships, and 5) Essential experimental and data processing protocols, 6) Metadata appropriate to the datasets so that they can be linked. Similar protocols will be followed for all transcriptomic, epigenetic, and genomic data that is generated in the course of this project. The data formats of fastq files, BAM, VCF, csv or tsv files, and code are standard across data repositories that store sequencing data.

Experimental metadata describing the sample collection and processing information, as well as relevant clinical information will be included as part of the data deposition process, format depends on the final NIH storage repository platform. This metadata will include, sample collection date, RNA integrity number (RIN) and any other relevant information. The [GEO](https://www.ncbi.nlm.nih.gov/geo/info/submission.html) submission procedure is aligned to the [MINSEQE](https://www.ncbi.nlm.nih.gov/geo/info/MIAME.html) (Minimum Information About a Next-generation Sequencing Experiment) guidelines which outline the minimum information that should be included when describing a sequencing study. [SRA](https://www.ncbi.nlm.nih.gov/sra/docs/submitmeta/) will provide internationally recognized project, study, sample, and experimental accession identifiers upon data submission to the repository.

**Element 4: 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)**

Relevant repositories for these data: GEO, dbGAP, SRA

**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.**

The scientific data will be findable and identifiable via a persistent unique identifier (such as a DOI or accession number) assigned by the chosen repository(ies). Additionally, the data will be indexed using standard indexing tools such as the repository's own search functionality and any relevant ontologies or controlled vocabularies.

**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 scientific data will be made available to other users no later than the time of an associated publication or end of the performance period, whichever comes first. The data will be available for a <insert timeframe e.g., minimum of 5 years from the date of publication> to ensure that the data remains accessible for future research and replication.

**Element 5: 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.**

Restrictions on subsequent access, distribution, or reuse of scientific data from this project include confidentiality protections. For example, any data that is proprietary or subject to non-disclosure agreements will not be shared without proper legal clearance. Informed consent will be obtained from all participants and all data will be de-identified to protect participant identities and maintain confidentiality. Additionally, necessary institutional review board approvals will be obtained prior to data collection and 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).**

Any data shared will be accompanied by a license agreement outlining the terms and conditions for reuse, including but not limited to the requirement to maintain participant confidentiality and obtain additional institutional review board approvals as necessary for any secondary research use.

**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).**

To protect participant privacy and confidentiality, shared data will be de-identified using the <specify de-identification method>. Note any other applicable laws or policies such as HIPAA. Additionally, as this research is focused on understanding the <insert scope of research>, any future use of the data must align with these research areas and not be used for any other purposes without obtaining additional informed consent from participants.

**Element 6: Oversight of Data Management and Sharing**

Describe how compliance with this Plan will be monitored and managed, frequency of oversight, and by
w Oversight of this plan will be managed by the <Project Investigator; ORCID, email >, and the <Research team> **(If using Genomics Research Center for data storage and public deposition, please include that here).** This will include regular reviews of this data management and sharing plan, as well as ensuring that all necessary approvals and consents are obtained and maintained. A point of contact will be provided for the <name of lab setting>, which will be responsible for all data management and sharing, including monitoring adherence <provide timeline>.

Data will be submitted upon specific data freeze milestones, e.g., these milestones and target timelines include:

End of 1st quarter: Data from the first quarter of the project will be reviewed and any necessary quality control measures will be taken.

End of 2nd quarter: All data from the second quarter of the project will be de-identified, if necessary, and prepared for submission to the chosen repository(ies).

End of 3rd quarter: Data from the third quarter of the project will be submitted to the chosen repository(ies) and made available to other researchers.

End of 4th quarter: Final data cleaning and curation will be performed, and any remaining data will be submitted to the chosen repository(ies) and made publicly available.

\*The data milestones are subject to change depending on the specific research project, but they will be reviewed and updated as necessary to ensure that data is being managed and shared in an efficient and timely manner at your institution (e.g., titles, roles).