The sample Data Management and Sharing Plan below is for a proposal conducting genomic research of human participants. It is one of four examples provided by NIDDK.

NIDDK Example Data Management and Sharing Plan – Genomic

Element 1: Data Type:

A. Types and amount of scientific data expected to be generated in the project:

This project will produce the following data types and datasets on 36 subjects totaling approximately 10 GB in size. The following final analysis data files will be produced:

- 4 phenotype/clinical datasets, 20 variables, 2 molecular datasets
- Molecular data types: Whole transcriptome shotgun sequencing (RNA-seq), Whole exome sequencing (WES), and methylome sequencing.

We propose to use Illumina's Real Time Analysis (RTA) software for the primary processing of sequencing images and the hclust algorithm with the ward method ("corrplot" R package) for the correlation analysis. We propose to use CASAVA 1.8.2 software to de-multiplex samples and generate raw reads and respective quality scores for DNA methylation analysis, BSMAP to compute the percent methylation scores and average conversion rates, the heatmap2 function of the "gplots" R package to generate the methylation heatmap, and the "stats" R package and the hclust algorithm to analyze the data dimensionality. We propose to use the "DMRcate" R package to determine statistically significant, differentially methylated CpG dinucleotides.

B. Scientific data that will be preserved and shared, and the rationale for doing so:

All phenotype/clinical datasets and molecular datasets will be preserved and shared to report the beta cell regeneration for diabetes via integration of molecular landscapes and aberrant methylation underlying insulin gene expression in human insulinomas.

C. Metadata, other relevant data, and associated documentation:

- **Study-level:** Study description; inclusion and exclusion criteria; history; selected publications; diseases/traits related to study (Medical Subject Headings [MeSH] terms); links to related genes; links to related resources
- **Dataset-level:** Dataset name, description, summary
- **Variable-level:** Variable name, description, statistical summary
- **Documentation:** Study protocols, data collection instruments/forms, data dictionary

Element 2: Related Tools, Software, and/or Code:

As described in section A, we will use open-source software for our data analysis:

- Pathway enrichment analysis - ClueGo, Cytoscape CluePedia, R package goseq
- Allele-specific expression analysis - R ggbio, Granges, ggplot2
- Co-expression analysis - R WGCNA
- Differential expression analysis - R limma

Element 3: Standards:

For key disease and traits, genes, and molecular data, we will use the following community standards:

- National Library of Medicine–hosted MeSH terms for diseases/traits
- National Human Genome Research Institute–supported Gene Ontology Resource for genes
- WES_markerset_grc37 and target_markerset_grc37 for WES markers
- Whole-genome bisulfite sequencing (WGBS) pipeline as part of the ENCODE Uniform Processing Pipelines

Type	Source	Platform	Comment
Whole Exome Sequencing	Illumina	HiSeq 2500	Insulinoma (tumor) and Paired Host Genome
RNA Sequencing	Illumina	HiSeq 2500	Insulinoma and FACS-sorted normal human beta cells
Targeted Chromosome 11 CpG Methylome Seq	Illumina	MiSeq	Targeted CpG methylome bisulfite seq of 1.35 Mbp in 11p15.5-15.4

Element 4: Data Preservation, Access, and Associated Timelines

A. Repository where scientific data and metadata will be archived:

We propose to submit our genomics data to database of Genotypes and Phenotypes (dbGaP) and mutation data to Catalogue of Somatic Mutations in Cancer (COSMIC).

B. How scientific data will be findable and identifiable:

Our proposed plan is to submit molecular and phenotypic data to the following public databases:

- DNAseq, RNAseq, and CpG Bisulfite seq raw data can be found at dbGaP (Accession number: phs001422.v1.p1)
- Mutation data can be found at COSMIC (ID number: COSP44132)

We will publish our findings in peer-reviewed journals. The repositories and journal(s) will provide metadata, persistent identifiers, and long-term access for both open and controlled access.

C. When and how long the scientific data will be made available:

Data will be made available as soon as possible or at time of publication, whichever comes first, and for at least 5 years' duration. Raw data, intermediate data, and the code/software/tools used to develop the published or submitted dataset will be shared at the time of data submission or publication and for at least 5 years' duration. All other scientific data will be made available no later than the end of the award.

Element 5: Access, Distribution, or Reuse Considerations

A. Factors affecting subsequent access, distribution, or reuse of scientific data:

The study datasets will be collected with the following informed consent:

- **Health/Medical/Biomedical (HMB) -** The dataset can only be used for studying health, medical or biomedical conditions and does not include the study of population origins or ancestry.

B. Whether access to scientific data will be controlled:

To maximize the appropriate sharing of scientific data and protect research participants' privacy and confidentiality, reuse of this dataset should use the following Data Use Limitations (DULs) under Controlled Access that is made available by a data repository only after approval.

- **Health/Medical/Biomedical (HMB) -** The dataset can only be used for studying health, medical or biomedical conditions and does not include the study of population origins or ancestry.
- **Institutional Review Board (IRB) Approval Required -** The requesting institution's IRB or equivalent body must approve the requested use.

C. Protections for privacy, rights, and confidentiality of human research participants:

To protect research participants' privacy and confidentiality, data submitted to the repositories does not include personally identifiable information such as names or addresses, and participants will be identified only by a unique identification number. Additional protections, such the approach for managing Health Insurance Portability and Accountability Act identifiers, including dates and geographical locations, will be used for de-identification. For example, dates are commonly obscured by replacing them with days from a reference event (e.g., enrollment).

Element 6: Oversight of Data Management and Sharing:

The Principal Investigator for this project, Dr. ABC, will ensure that this Data Management and Sharing Plan is followed. The institutional official (title and role), will be responsible for oversight of compliance with the accepted Data Management and Sharing (DMS) Plan. Compliance will be evaluated annually during the award period and progress towards the plan's DMS activities will be included in the annual Research Performance Progress Report (RPPR) submitted to the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) Project Officer. At the project conclusion, the final progress report will summarize how the DMS objectives were fulfilled and provide links to the shared dataset(s).