Data Management Plan (DMP)

Project Title: The Interplay of Circulating MicroRNAs and Gut Microbiome Diversity in Predicting Early-Onset Type 2 Diabetes Mellitus Risk

Date: October 2025
Version: 1.0

1. Data Collection and Creation

1.1 Data Types:
The study will generate three primary types of data:

1. **Clinical Data:** Demographics, medical history, anthropometrics (BMI, waist circumference), lab values (HbA1c, glucose, lipids).
2. **Sequencing Data (Raw):** Next-Generation Sequencing (NGS) fastq files for microRNA and 16S rRNA/Shotgun Metagenomics.
3. **Processed/Analyzed Data:** Differentially expressed miRNA lists, microbiome diversity metrics, taxonomic abundance tables, and final predictive model parameters (e.g., Random Forest outputs).

1.2 Data Capture:
Clinical data will be entered directly into a secure, encrypted, and validated Electronic Data Capture (EDC) system (e.g., REDCap). Sequencing data will be generated by the core laboratory facility and transferred via secure, password-protected institutional cloud storage.

2. Documentation and Metadata

2.1 Metadata Standard:
Metadata for sequencing data will adhere to the Minimum Information about a 16S rRNA Gene Sequence (MIxS) standards for microbiome data and the Minimum Information about a microRNA Experiment (MIAME) standards for small RNA data.
2.2 Documentation:
A detailed data dictionary will be maintained for all clinical variables, including variable names, definitions, allowed values, and units of measure. All bioinformatic scripts (e.g., for processing fastq files) will be version-controlled and stored in a private institutional repository (e.g., Git).

3. Storage and Backup

3.1 Active Data Storage:
Clinical data will be stored on the secure institutional server, backed up nightly. Raw sequencing files (large data volume) will be stored on a high-capacity, dedicated research computing cluster with built-in redundancy (RAID 6).
3.2 Backup Strategy:
A daily backup of all clinical and processed data will be performed, with monthly backups archived offline. The retention period for backups will be 5 years.

4. Ethics, Legal Compliance, and Data Security

4.1 Privacy and Anonymization:
All identifying information will be removed immediately after the initial data capture. Data used for analysis will be linked only by a Study-Specific Alphanumeric Code (SAAC). The master key linking SAACs to PHI will be stored in a separate, locked physical cabinet and a separate, encrypted digital file, accessible only by the PI and Co-PI.
4.2 Ownership:
The institutional entity (e.g., University/Hospital) will be the data steward. The PI will hold the primary responsibility for data management and security.

5. Data Sharing and Preservation

5.1 Data Sharing Policy:
De-identified, raw, and processed sequencing data will be made publicly available no later than the publication date of the main findings manuscript.

- **Microbiome Data:** Submitted to the NCBI **Sequence Read Archive (SRA)**.
- **microRNA Expression Data:** Submitted to the NCBI **Gene Expression Omnibus (GEO)**.

5.2 Preservation:
Data will be preserved for a minimum of 10 years following project completion. Preservation will utilize institutional long-term archival storage solutions. Clinical data will be shared only in aggregated or de-identified forms to protect participant privacy.