**DATA MANAGEMENT PLAN**

**Project Title: FINDING, FOLDING, AND ELUCIDATING FUNCTIONS OF RNA STRUCTURES IN THREE VIRAL TARGETS OF CLINICAL IMPORTANCE.**

**This project investigates the intricate RNA structures found within three viruses of significant clinical interest. Understanding RNA structure and function is essential for developing insights into viral pathogenesis and for designing potential RNA-based therapeutics. This investigation will combine techniques such as chemical probing, bioinformatics analysis, and cryo-EM to uncover RNA structural motifs, characterize their interactions with proteins, and determine high-resolution 3D structures. Effective data management is crucial for the success of this project. The following plan outlines the strategies for data generation, storage, preservation, accessibility, and security.**

**Data Generation and Types**

**This project will produce a diverse set of data, including:**

1. **FastQ files: Raw RNA sequencing reads generated from chemical probing experiments.**
2. **Processed sequencing data: This includes alignment files, reactivity profiles, and the results of differential analyses.**
3. **Structural models: We will develop both 2D secondary RNA structure predictions and 3D tertiary structures (derived from cryo-EM or modeling).**
4. **Cryo-EM datasets: 2D projection images and 3D reconstructions visualize important RNA complexes.**
5. **Analysis Scripts: Custom code will be created for all stages of data processing, structure prediction, and analysis.**

**Data Storage and Preservation**

**FastQC: the project involves the generating of Fastq files resulting from the chemical probing of RNA and subsequent sequencing using Illumina technology (iSeq). These files containing raw sequences are analyzed to determine the reactivity profile before being used in the scanfold program. To accommodate the large data set that will be generated from this program, the team has been allocated a dedicated space on the Iowa State University High-Performance Computing (HPC) system. By centralizing the storage of Fastq files within the lab's designated area on the HPC cluster, accessibility, security, and scalability are ensured. This strategic decision facilitates seamless access to the primary data for downstream analyses while safeguarding against data loss or corruption.**

**Cryo-EM data: During the course of the project, structures of significant importance will visualize using the Cro EM facility in the institute. The availability of Cryo-EM visualization facilities at Iowa State University further enriches the research environment, facilitating real-time exploration and interpretation of complex structural data. These high-resolution structural data provide invaluable insights into the spatial arrangement and conformational dynamics of selected RNA structures, shedding light on their functional roles and interactions. To manage the voluminous Cryo-EM datasets effectively, the images will be stored on encrypted hard drive for easy access and on the HPC cluster alongside the Fastq files. By consolidating all research data within a unified infrastructure, synergy is fostered, enabling interdisciplinary collaborations and holistic data-driven investigations.**

**Analysis Script: the development of analysis scripts constitutes an integral component of the research workflow. These scripts, tailored to specific analytical tasks and computational algorithms, serve as the backbone of data processing and interpretation will be saved on the lab github page,** [moss-lab (github.com)](https://github.com/moss-lab)**. A decentralizable approach not only promotes transparency and reproducibility but also fosters knowledge dissemination and community engagement.**

**Laboratory/Assay Result: A version control of all lab result (qPCR, Dual Luciferase assay reporter, gel images, PAGE etc), and cell line passages will be documented and saved on cybox using the version control system and that is accessible to all laboratory members. This will ensure that the data generated are updated regularly.**

**Accessibility and Sharing**

**Secure shared access will be established within our research team. To promote scientific discovery, we plan to share our findings with the broader research community. Raw FastQ files will be deposited in public repositories like the Sequence Read Archive (SRA). Structural models will be submitted to the Protein Data Bank (PDB) or other relevant databases. Analysis scripts will be made publicly available on platforms like GitHub.**

**Data Security**

**We will adhere to all security protocols established by the HPC. Any sensitive data will be handled according to specific ethical guidelines. Primary storage will utilize Iowa State University's High-Performance Computing (HPC) cluster, with space allocated specifically for our lab. Data will be meticulously organized using clear directory structures, descriptive naming conventions, and appropriate metadata standards. Version control will be implemented for scripts and analysis pipelines, ensuring reproducibility. Regular backups to secondary locations will provide redundancy and safeguard against data loss.**

**In conclusion, effective data management is the very important to achieve a successful and reproducible research, thus, by implementing robust strategies for handling diverse data types, including Fastq files, Cryo-EM structures, and analysis scripts, the I will ensure that the integrity, accessibility, and scalability of research data is been maintained. Through the synergistic integration of computational resources, visualization tools, and collaborative platforms, such as the Iowa State University HPC cluster and GitHub repository, the research project exemplifies a holistic approach to data-driven discovery. By embracing these principles of data management excellence, the research team advances scientific knowledge and contributes to the collective quest for biomedical innovation and societal impact.**