# Element 1: Data Type

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

A minimum of 202 participants will be enrolled in the study. Clinical data will be collected through clinical coordinators and computer-assisted personal interviews and direct participant survey. Clinical data include demographic data, medical history, medications, lab tests performed in a central laboratory, among other data pertinent to the study. Research laboratory data include results of HLA-subtyping and immunoassays from biospecimens.

Clinical data will be captured using the REDCap electronic data management (EDM) system at the University of Arizona. REDCap software support is provided by the University of Arizona Center for Biomedical Informatics, and administration of REDCap database and web server is managed through limited intranet access from three consoles to the interface Virtual Server located at University of Arizona Information Technology Services.

The University of Arizona Genomics Core (UAGC) will be responsible for HLA-haplotyping and scRNA-Seq data generation. The UAGC is a CAP/CLIA certified lab and as such all procedures have detailed and versioned tracked SOPs. Data is stored securely using a laboratory information management system (LIMS) and sample provenance and chain of custody is clearly tracked and documented. Samples are also securely stored with barcoded sample tracking information which don’t contain any participant identifying information.

CD8 and CD4 T cell data, utilizing highly multiplexed spectral flow cytometry, will be conducted per lab SOPs, sample data are stored on a secured computer and metadata don’t contain any participant identifying information.

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

Clinical data that will be preserved and shared are demographic data, medical history, lab tests performed by the clinical site and central laboratory, and physical exams data, among other data pertinent to the study. Research laboratory data that will be preserved and shared include results of cytokine analysis and other immunoassays from biospecimens. Genomic data that will be preserved include raw variant call format (.vcf) and flowcytometry standard file (.fsc). The raw files will be stored on secured (HIPAA for clinical and password protected for de-identified original/raw data) servers for 10 years after study completion.

De-identified and raw data will be archived at UArizona on the CyVerse (NSF-funded cloud-based infrastructure) and will be available to all registered users. Any de-identified data shared with registered users will be shared through the OAIC Knowledge and Data Dashboard housed on CyVerse and will require secure web authentication, data logging, and Secure Sockets Layer (SSL) encryption. Findability, Accessibility, Interoperability, and Reusability (FAIR) standards will be followed for all bioinformatic and statistical workflows. All protocols for specimen procurement, storage, handling, and processing (genomic, biometric, etc.) as well as analytic workflows will be made available to the scientific community. This application will use a protocol template (protocol.io or similar) and allow registered users to examine changes through a version-controlled repository. Any updated specimen processing protocols will be approved by the executive committee and updated versions will be uploaded to the repository immediately after approval. Specimen metadata (e.g., participant ID, batch, protocol version number, sequencer ID), storage and handling procedures, and sample QC metrics will also be shared with pre-processed (raw) data and processing protocols.

De-identified data, along with computational and statistical reproducible files will be shared on ImmPORT upon manuscript publication. Additionally, raw HLA typing data (.vcf) will be shared to NCBI SRA, and raw spectral flow data (.fsc) will be shared to FlowRespository within six months of study completion.

# Metadata, other relevant data, and associated documentation:

The protocol, sample informed consent, case report forms, data dictionary, and code book will be made accessible in data repositories where data are shared. For data submitted to ImmPORT, variable-level metadata will be provided using a templated data dictionary, and will include details of Common Data Elements, definitions, and standards used for data collection and sharing.

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

Clinical and laboratory data will be collected in the electronic data capture system (REDCap) and any data sharing will use application programming interfaces (APIs) that will provide seamless and automated integration of de-identified data for analyses. For genomic and immunologic data analysis, containerized (Docker) open- source workflows using markdown or web-based computing (e.g., Jupyter Notebooks) will be used for both bioinformatics and statistical analyses. GitHub will be used to share data and analyses with ImmPORT so that versioning can be maintained.

# Element 3: Standards:

Shared data will be de-identified, and archived files (including code) will be stored on secured server(s) for at least 10 years after study completion.

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

1. **Repository where scientific data and metadata will be archived:**

ImmPORT (Clinical and laboratory data); Sequence Read Archive (HLA-typing); FlowRepository (spectral flow cytometry data).

# How scientific data will be findable and identifiable:

Clinical and laboratory data will be findable and identifiable using DOI. SRA will be identifiable using sequence record accession numbers generated by NCBI.

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

ImmPORT study data and reproducible code will be available at manuscript acceptance. SRA and FlowRepository data will be shared no more than six months upon study completion. Data will be preserved within the repositories for at least three years following the completion of the grant, as required by federal retention guidelines.

Data sharing of de-identified data, SOPs, and code can also be shared outside repositories at the discretion of the project leadership.

# Element 5: Access, Distribution, or Reuse Considerations

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

CyVerse Data Commons services will be used to manage, organize, preserve, publish, discover, and reuse data that will be shared with registered users. When possible, a permanent link will be established from the CyVerse Cyberinfrastructure to allow direct publishing of data to external repositories through the Data Commons portal. De-identified data will be shared when needed in accordance with applicable state and federal laws, regulations and rulings related to the protection of human subjects.

# Whether access to scientific data will be controlled:

Data will be controlled access with the General Research Use Data Use Limitation, as allowed by the informed consent and the institutional certification*.*

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

Informed consent documents used for the proposed clinical trial will include explicit language informing the participant or legally authorized representative that residual biological specimens, including HLA subtypes, may be stored in a biorepository for other scientific investigations. Privacy and confidentiality protections consistent with applicable federal, Tribal, state, and local laws, regulations, and policies will be followed. Data will be de-identified by removing all 18 HIPAA identifiers prior to sharing, and the study will have a Certificate of Confidentiality from NIH.

# Element 6: Oversight of Data Management and Sharing:

Data will be submitted by a data administrator from the project team. The data administrator will oversee data collection, analysis, storage, and sharing. Compliance with the plan will be monitored by the leadership team routinely. The leadership team will ensure data are submitted and shared according to this Data Management and Sharing Plan.