

**Institute of Primate Research**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SOP No.** | **Issue Number** | **Issue Date** | **Revision Status** | **Revision Date** |
| **SOP/KIPRE/RPD/DSAS/3.1.76** | **Version 01** | **October 2025** | **-** | **-** |

**STANDARD OPERATING PROCEDURE (SOP) DOCUMENT**

**Genome and Proteome Data Management**

|  |  |  |  |
| --- | --- | --- | --- |
| **Approvals** |  |  |  |
|  | **Name** | **Signature** | **Date** |
| **Developed by:** | \_Patrick Waweru Mwaura\_ | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **\_6th October; 2025\_** |
|  | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** |
|  | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** |
| **Reviewed by:** | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** |
| **Approved by:** | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

**Table of Contents**

[1. PURPOSE 4](#_Toc144316958)

[2. SCOPE 4](#_Toc144316959)

[3. PERSONS RESPONSIBLE: 4](#_Toc144316960)

[4. FREQUENCY 4](#_Toc144316961)

[5. MATERIALS 5](#_Toc144316962)

[6. PROCEDURE 5](#_Toc144316963)

[7. REFERENCES 5](#_Toc144316964)

# 

# PURPOSE

To establish standardized procedures for the **secure, compliant, and reproducible management of genomic and proteomic datasets** within DS&AS, ensuring that all data are:

* Handled according to institutional policies and legal requirements (e.g., Kenya Data Protection Act 2019, GDPR).
* Stored, processed, and shared in alignment with **SOP 6 (Data Access and Authentication)**, **SOP 7 (Data Storage, Backup, and Disaster Recovery)**, and **SOP 9 (Data Sharing and Anonymisation)**.
* Annotated and structured to support reproducible research and interoperability in line with **FAIR principles**.

# SCOPE

Covers all genomic and proteomic datasets managed by DS&AS, including:

* Raw sequencing and mass-spectrometry data (FASTQ, BAM, FASTA, RAW).
* Processed and annotated datasets (VCF, GTF, protein expression tables).
* Associated metadata describing samples, experimental conditions, and analytical workflows.
* Activities related to storage, versioning, analysis, and secure sharing in accordance with **SOPs 6, 7, and 9**.

# PERSONS RESPONSIBLE:

* **Bioinformatician / Data Scientist:** Oversees genomic and proteomic data preprocessing, quality control, annotation, and reproducible analysis pipelines (**linked to SOPs 3, 4, and 5**).
* **Data Engineer:** Implements and maintains secure databases, version control, backups, and access management (**linked to SOPs 6, 7, and 8**).
* **Head of DS&AS:** Ensures overall compliance with institutional policies, national regulations, and international standards (**linked to SOPs 1, 2, and 9**).
* **Principal Investigator (PI):** Provides experimental design, sample metadata, and ensures alignment of project data with approved protocols.
* **Data Protection Officer (DPO):** Reviews access, sharing, and anonymisation to ensure regulatory compliance.

# FREQUENCY

* **Continuous:** Data management, preprocessing, and access control are performed throughout the project lifecycle (**aligned with SOPs 6, 7, 8**).
* **Annual Audits:** Comprehensive review of data integrity, storage, access, and compliance with regulatory and institutional standards (**linked to SOPs 7 and 9**).
* **Triggered Reviews:** Additional audits or updates occur whenever regulatory changes, major protocol amendments, or security incidents arise.

# MATERIALS

* **Secure Storage & Computing:** Encrypted on-premise servers, cloud storage (AWS, Azure), and version-control systems (**linked to SOPs 6 and 7**).
* **Reference Databases:** Public genomic/proteomic resources such as GenBank, Ensembl, UniProt, and proteomics repositories.
* **Metadata Standards:** Templates adhering to **MIAME (Minimum Information About a Microarray Experiment)** and **MIAPE (Minimum Information About a Proteomics Experiment)** to ensure reproducibility (**linked to SOP 8**).
* **Data Management Policies:** Institutional Data Protection and Sharing Policy, including anonymisation and access guidelines (**linked to SOPs 1, 2, and 9**).
* **Analysis Tools:** Bioinformatics software and pipelines (e.g., R, Python, Galaxy, Nextflow, Snakemake).
* **Documentation Templates:** Standardized forms for data dictionaries, dummy tables, and version-controlled workflow records (**linked to SOP 4**).

# PROCEDURE

1. **Data Collection & Storage:**

* Store raw genomic and proteomic data in secure servers or cloud repositories immediately after generation (**SOPs 6 & 7**).
* Assign project-specific identifiers and record storage location in the data registry (**SOP 8**).

1. **Metadata Capture:**

* Document experimental details, sample information, and processing steps using **MIAME/MIAPE-compliant templates** (**SOPs 3 & 8**).
* Link metadata to datasets to support reproducibility and FAIR principles (**SOP 1**).

1. **Quality Control:**

* Perform sequence or proteome QC using standardized tools (e.g., FastQC, ProteoQC, or equivalent pipelines).
* Document QC outcomes and any corrective actions in the project repository (**SOP 4**).

1. **Access Control:**

* Implement role-based access for all users according to data sensitivity (**SOP 6**).
* Log all access and changes for audit purposes (**SOP 9**).

1. **Archiving & Backup:**

* Maintain incremental and full backups with version-controlled archives (**SOP 7**).
* Ensure offsite/cloud mirrors for disaster recovery.

1. **Data Sharing & Compliance:**

* Anonymise or pseudonymise human-derived data before sharing externally (**SOP 9**).
* Only release datasets with formal approvals from the Head of DS&AS and DPO.

1. **Documentation & Reporting:**

* Maintain detailed records of all steps, QC results, and version history for audit and reproducibility (**SOPs 4 & 5**).

# REFERENCES

1. Kenya Data Protection Act, 2019.
2. General Data Protection Regulation (GDPR), Regulation (EU) 2016/679.
3. FAIR Data Principles: Findable, Accessible, Interoperable, Reusable.
4. MIAME: Minimum Information About a Microarray Experiment.
5. MIAPE: Minimum Information About a Proteomics Experiment.
6. SOP 1: Policies and Strategies for DS&AS.
7. SOP 2: Alignment with Institutional and National Regulations.
8. SOP 4: Statistical Analysis Plans (SAPs).
9. SOP 6: Data Access and Authentication Procedures.
10. SOP 7: Data Storage, Backup, Encryption, and Disaster Recovery.
11. SOP 8: Database and Workflow Management.
12. SOP 9: Data Sharing, Anonymisation, and Compliance.

# ****APPENDIX / FORMS****

**A. Data Management Forms & Templates**

* **Genome/Proteome Data Dictionary Template:** Captures dataset variables, units, and descriptions.
* **QC Log Sheet:** Tracks quality control outcomes (e.g., sequence quality, coverage, proteomics metrics).
* **Dummy Tables & Figures Template:** For pre-specifying tables and figures in analysis.
* **Version Control & Audit Log Form:** Records dataset versions, backup dates, and access changes.
* **Metadata Capture Template:** MIAME/MIAPE-compliant template for experimental details.
* **Data Sharing Approval Form:** For external release requests, including DPO and Head of DS&AS sign-off.
* **Access Request Form:** Requests for role-based dataset access within DS&AS.

**B. Standard Operating Guidelines References**

* Links to SOPs 1–9 for cross-referenced procedures in policy compliance, access control, storage, backup, and sharing.