Institute of Primate Research

STANDARD OPERATING PROCEDURE (SOP) DOCUMENT

Bioinformatics pipelines (from raw sequence data to analysis)

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

Approvals

	Name	**Signature**	**Date**
Developed by:	_Patrick Waweru Mwaura_		**_6th October; 2025_**
Reviewed by:			
Approved by:			

Table of Contents

1. PURPOSE

To establish standardized workflows for the design, implementation, and execution of bioinformatics pipelines, ensuring reproducibility, accuracy, and efficiency while maintaining compliance with institutional policies (SOPs 1, 4, 7, 8, 12), data protection regulations (SOP 6, 9), and best practices in genomic and proteomic data analysis.

2. SCOPE

Applies to all DS&AS-supported projects that involve bioinformatics analyses of genomic and proteomic data, from raw sequence acquisition through processing, quality control, alignment, annotation, and downstream statistical or functional analysis. Includes pipelines implemented on local servers, HPC clusters, or cloud platforms.

3. PERSONS RESPONSIBLE:

- **Bioinformatician:** Designs, implements, and executes bioinformatics pipelines; ensures reproducibility and accuracy of analyses.
- **Computational Biologist / Data Scientist:** Interprets pipeline outputs, validates findings, and provides feedback for workflow optimization.
- **Head of DS&AS:** Reviews and approves all pipeline workflows before deployment; ensures compliance with institutional policies, data governance standards, and SOPs 4, 7, 8, and 12.

4. FREQUENCY

- **Initial Validation:** Each bioinformatics pipeline must undergo validation prior to its first deployment on project data.
- **Periodic Updates:** Pipelines are reviewed and updated whenever new tools, algorithms, reference genomes/builds, or regulatory/data governance requirements are introduced.
- **Re-Validation:** Any major pipeline update triggers a full re-validation to ensure continued accuracy, reproducibility, and compliance with institutional and regulatory standards.

5. MATERIALS

- **Workflow Management Systems:** Nextflow, Snakemake, Galaxy for pipeline orchestration and reproducibility.
- **Analysis Tools:** Alignment, variant calling, and functional analysis software (e.g., BWA, GATK, DESeq2, BLAST, HISAT2).
- **Computational Infrastructure:** High-performance computing (HPC) clusters, cloud platforms, or local servers.
- **Version Control:** Git or other repository systems for workflow versioning and collaborative development.
- **Reference Data:** Genome builds, annotation files, proteome references, and relevant metadata standards (e.g., MIAME/MIAPE).
- **Documentation Templates:** SOP-linked pipeline documentation, logging, and reporting templates.

6. PROCEDURE

1. **Pipeline Design:** Define the full workflow, including data quality control (QC), preprocessing, alignment/mapping, variant calling, annotation, and downstream analyses.
2. **Implementation:** Develop reproducible pipelines using workflow management systems (Nextflow, Snakemake, Galaxy), ensuring modularity and portability.
3. **Testing & Validation:** Validate pipelines on benchmark datasets to confirm accuracy, reproducibility, and compliance with SOP 4 (SAPs) and SOP 12 (genome/proteome data management).
4. **Execution:** Run pipelines on designated HPC or cloud infrastructure; systematically log all outputs, errors, and runtime metrics.
5. **Version Control:** Maintain all pipeline scripts, configuration files, and container images (if applicable) in Git repositories or institutional version-control systems.
6. **Documentation & Archiving:** Document pipeline design, parameters, reference data, and results; archive all materials in the DS&AS repository for transparency and reproducibility.
7. **Updates & Re-Validation:** Upon tool or reference data updates, revise pipelines, document changes, and re-validate to maintain compliance and accuracy.

7. REFERENCES

1. Afgan, E., et al. (2018). *The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update*. Nucleic Acids Research, 46(W1), W537–W544.
2. Köster, J., & Rahmann, S. (2012). *Snakemake—a scalable bioinformatics workflow engine*. Bioinformatics, 28(19), 2520–2522.
3. Di Tommaso, P., et al. (2017). *Nextflow enables reproducible computational workflows*. Nature Biotechnology, 35, 316–319.
4. Li, H., & Durbin, R. (2009). *Fast and accurate short read alignment with Burrows–Wheeler transform*. Bioinformatics, 25(14), 1754–1760.
5. Van der Auwera, G.A., & O'Connor, B.D. (2020). *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. O'Reilly Media.
6. MIAME (Minimum Information About a Microarray Experiment) & MIAPE (Minimum Information About a Proteomics Experiment) Guidelines.
7. KIPRE Institutional SOPs 4, 7, 8, 12 (SAPs, Data Storage, Database Management, Genome/Proteome Data Management).

8. APPENDIX

Appendix A – Pipeline Documentation Template

- Pipeline name and version
- Workflow diagram and steps
- Input data types and formats
- Reference genomes/annotation used
- Software tools and versions
- Parameter settings and rationale
- Output description
- Validation dataset and results
- Responsible personnel

Appendix B – Pipeline Update & Re-Validation Log

- Date of update
- Nature of change (tool, parameter, reference data)
- Validation results after update

- Approving personnel

Appendix C – Data Access and Compliance Checklist

- Access level required
- Data anonymisation/pseudonymisation applied
- Regulatory compliance verified (SOP 9, DPA 2019, GDPR if applicable)
- Archiving and version control confirmation