

**DEVELOPMENT OF A COMPUTATIONAL PIPELINE FOR NEXT-
GENERATION SEQUENCING DATA ANALYSES USING
NEXTFLOW AND DOCKER**

**OWOLABI, PAUL JESUSANMI
(21PBF02261)**

B.Sc Microbiology, Obafemi Awolowo University, Ile-Ife, Nigeria

AUGUST, 2023

**DEVELOPMENT OF A COMPUTATIONAL PIPELINE FOR NEXT-
GENERATION SEQUENCING DATA ANALYSES USING
NEXTFLOW AND DOCKER**

BY

**OWOLABI, PAUL JESUSANMI
(21PBF02261)**

B.Sc Microbiology, Obafemi Awolowo University, Ile-Ife, Nigeria

**A DISSERTATION SUBMITTED TO THE SCHOOL OF
POSTGRADUATE STUDIES, IN PARTIAL FULFILMENT OF THE
REQUIREMENTS FOR THE AWARD OF A MASTER OF SCIENCE
DEGREE IN BIOINFORMATICS IN THE DEPARTMENT OF
COMPUTER AND INFORMATION SCIENCES, COLLEGE OF
SCIENCE AND TECHNOLOGY, COVENANT UNIVERSITY, OTA,
OGUN STATE, NIGERIA**

AUGUST, 2023

ACCEPTANCE

This is to attest that this dissertation has been accepted in partial fulfilment of the requirements for the award of the degree of Master of Science in Bioinformatics in the Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota, Nigeria.

Miss Adefunke F. Oyinloye
(Secretary, School of Postgraduate Studies)

Signature and Date

Prof. Akan B. Williams
(Dean, School of Postgraduate Studies)

Signature and Date

DECLARATION

I, **OWOLABI, PAUL JESUSANMI (21PBF02261)** hereby declare that this dissertation titled “**DEVELOPMENT OF A COMPUTATIONAL PIPELINE FOR NEXT-GENERATION SEQUENCING DATA ANALYSES USING NEXTFLOW AND DOCKER**” is a representation of my work and is written and implemented by me under the supervision of Dr. Itunuoluwa M. Isewon of the Department of Computer and Information Sciences, Covenant University, Ota, Nigeria. I attest that this dissertation has in no way been submitted either wholly or partially to any other university or institution of higher learning for the award of a masters’ degree. All information cited from published and unpublished literature has been duly referenced.

OWOLABI, PAUL JESUSANMI

Signature and Date

CERTIFICATION

This is to certify that this dissertation titled **“DEVELOPMENT OF A COMPUTATIONAL PIPELINE FOR NEXT-GENERATION SEQUENCING DATA ANALYSES USING NEXTFLOW AND DOCKER”**, is an original research carried out by **OWOLABI, PAUL JESUSANMI (21PBF02261)** and meets the requirements and regulations governing the award of Master of Science (M.Sc.) degree in Bioinformatics from the Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota, and is approved for its contribution to knowledge and literary presentation.

Dr. Itunuoluwa M. Isewon
(Supervisor)

Signature and Date

Prof. Olufunke O. Oladipupo
(Head of Department)

Signature and Date

Prof. Adebukola S. Onashoga
(External Examiner)

Signature and Date

Prof. Akan B. Williams
(Dean, School of Postgraduate Studies)

Signature and Date

DEDICATION

I am dedicating this work to the Almighty God, who is my true source, and to those committed to the furtherance of life science research.

ACKNOWLEDGEMENTS

I begin by expressing my heartfelt appreciation to my God for His grace bestowed unto me to embark upon and conclude this research project. I appreciate the Chancellor of Covenant University, Dr. David O. Oyedepo for his steadfast commitment to the institution's mission that has provided us with a strong foundation to excel in our pursuits. Thanks to the Vice Chancellor, Prof. Abiodun H. Adebayo and the Dean School of Postgraduate Studies, Prof. Akan B. Williams for their emphasis on creating an enabling environment that has empowered both faculty and students to thrive and contribute meaningfully to the academic community. I appreciate the Head of the Department of Computer and Information Sciences, Prof. Olufunke O. Oladipupo for her unflinching support for scholarly works in the department.

I appreciate my Supervisor, Dr. Itunuoluwa M. Isewon, who is also the Postgraduate Coordinator of the Department of Computer and Information Sciences. I equally want to thank Prof. Ezekiel F. Adebisi for his mentorship and Dr. Yagoub A. Adam for his immense help and great contribution towards the success of this project. My gratitude also goes to the entire faculty, staff, and students of the Department of Computer and Information Sciences for their support and guidance. I acknowledge the Covenant Applied Informatics and Communication Africa Centre of Excellence (CApIC-ACE) and the World Bank for providing me this platform by supporting my studentship.

I am indeed grateful to my parents Pastor Stephen Owolabi and Mrs. Deborah Owolabi for their immeasurable support. I thank my siblings John Owolabi, Glory Owolabi and God's Love Owolabi for being wonderful. I am grateful to Fesobi Oluwamuyiwa for his support in the early days of my research work. Finally, I wish to express my profound gratitude to my friends and colleagues especially Faith Adegoke, Jumoke Adeyemi, Erika Baiguerel, Stephen Binaansim, Samuel K T. Owusu-Ansah and Blessing Onyido for making this journey worthwhile.

TABLE OF CONTENTS

CONTENTS	PAGES
COVER PAGE	i
TITLE PAGE	ii
ACCEPTANCE	iii
DECLARATION	iv
CERTIFICATION	v
DEDICATION	vi
ACKNOWLEDGEMENTS	vii
TABLE OF CONTENTS	viii
LIST OF FIGURES	x
LIST OF TABLES	xii
LIST OF ABBREVIATIONS	xiii
ABSTRACT	xiv
CHAPTER ONE: INTRODUCTION	
1.1 Background of the Study	1
1.2 Statement of the Research Problem	8
1.3 Aim and Objectives	9
1.4 Significance of the Study	10
1.5 Scope of the Study	10
1.6 Organization of the Study	11
CHAPTER TWO: LITERATURE REVIEW	12
2.1 Preamble	12
2.2 Conceptual Review	12
2.2.1 Genomics and Personalized Medicine in Africa	12
2.2.2 Genome Alignment	14
2.2.3 Variant Calling, Filtration and Annotation	15
2.2.4 Estimating Genome Characteristics	16
2.2.5 Workflow Management Systems	17
2.3 Review of Related Works	20
CHAPTER THREE: METHODOLOGY	23
3.1 Preamble	23

3.2	Materials	23
	3.2.1 Genomic Sequence Retrieval	23
3.3	Methods	25
	3.3.1 Reference Preparation	25
	3.3.2 Sample Quality Control Check	25
	3.3.3 Mapping	25
	3.3.4 Genome Size Heterozygosity	26
	3.3.5 Variant Calling	26
	3.3.6 Variant Annotation	27
	3.3.7 Nextflow Scripting	27
	3.3.8 Containerisation	28
	3.3.9 Implementation	29
CHAPTER FOUR: RESULTS AND DISCUSSION		31
4.1	Preamble	31
4.2	Results	31
	4.2.1 Nextflow Execution Timeline and Report	32
	4.2.2 Resource Usage	33
	4.2.3 Sample Characteristics	39
	4.2.4 Outcome of the Analysis of the Sample Data	40
4.3	Discussion	53
CHAPTER FIVE: CONCLUSION AND RECOMMENDATIONS		55
5.1	Conclusion	55
5.2	Contribution to Knowledge	55
5.3	Recommendations	55
5.4	Limitations of the Study	56
5.5	Areas for Further Research	56
REFERENCES		57
APPENDIX A		72
APPENDIX B		80

LIST OF FIGURES

FIGURES	TITLE OF FIGURES	PAGES
2.1	Ultrarapid Genome Sequencing Pipeline	22
3.1	An overview of the WES data analysis process	24
3.2	GATK Variant Calling Best Practices	29
3.3	The implementation of the pipeline using Nextflow	30
4.1	Snapshot of the successful run of the main.nf Nextflow script in Ubuntu	31
4.2	Processes execution timeline	32
4.3	Nextflow workflow report	33
4.4	Overview of the CPU usage	34
4.5	Overview of the % Requested CPU usage	34
4.6	Overview of the Physical Memory usage	35
4.7	Overview of the Virtual Memory usage	36
4.8	Overview of the % Requested Physical Memory used	36
4.9	Overview of the Task execution real-time	37
4.10	Overview of the Number of bytes read by the input devices	38
4.11	Overview of the Number of bytes written by the output devices	38
4.12	Some metrics on each of the tasks in the workflow	39
4.13	Snapshot of the characteristics of the test data on SRA webpage	40
4.14	The sequence counts of the paired ends of the test data	41
4.15	The sequence duplication levels of the paired ends of the test data	41
4.16	The adapter content of the paired ends of the test data	42
4.17	Per base N content of the paired ends of the test data	42

4.18	The mean Quality scores of the paired ends of the test data	43
4.19	The per sequence GC content of the paired ends of the test data	43
4.20	The per sequence Quality scores of the paired ends of the test data	44
4.21	Status checks of the paired ends of the test data	45
4.22	Linear plot of the heterozygosity of the test data	46
4.23	Log plot of the heterozygosity of the test data	47
4.24	Transformed linear plot of the heterozygosity of the test data	48
4.25	Transformed log plot of the heterozygosity of the test data	49
4.26	General summary of the SnpEff results of the test data	50
4.27	Number of variant effects by type and region	52
4.28(a)	Base changes (SNPs) in the test data	53
4.28(b)	Total number of transitions and transversions in the test data as well as their ratio	53

LIST OF TABLES

TABLE	TITLE OF TABLES	PAGES
2.1	Features of some of the most common workflow management systems	19
4.1	The number of variants by type	51
4.2	The number of effects by impact	51
4.3	The number of effects by functional class	51

LIST OF ABBREVIATIONS

ADME	Absorption, Distribution, Metabolism, and Excretion
BWA	Burrows-Wheeler Aligner
CNVs	Copy Number Variations
CWL	Common Workflow Language
DNA	Deoxyribonucleic Acid
GATK	Genome Analysis Toolkit
GRCh38	Genome Reference Consortium Human Build 38
GWAS	Genome Wide Association Studies
HTS	High-throughput Sequencing
NCBI	National Centre for Biotechnology Information
NGS	Next generation Sequencing
SNPs	Single Nucleotide Polymorphisms
SNVs	Single Nucleotide Variants
SRA	Sequence Read Archive
SVs	Structural Variants
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing

ABSTRACT

Major advances in genomics studies, particularly the introduction of high-throughput sequencing and the evolution of genotyping platforms have led to the emergence of big data in the biological sciences and a growing need to make sense of this data. This has largely fostered the evolution of methods and tools for genomic data analysis (especially of diseased conditions) with the aim of uncovering the genotype-phenotype relationships in such diseased conditions. Due to the growing complexity and volume of next-generation sequencing data available in biological sciences, there is a growing need to developed pipelines that can handle these data while automating most of the steps involved in these analyses. The aim of this study is to develop a computational pipeline for the analysis of next-generation sequencing data using Nextflow and Docker. Since different steps and tools are involved in the analysis of whole genome and whole exome sequencing data, the aim of the study was achieved by developing scripts for selected genome analysis tools, building a computational pipeline for the selected tools and performing unit and integration testing for the pipeline. The pipeline which was built on the framework of the well-established GATK best-practices workflow, integrated the following tools: FastQC, MultiQC, Jellyfish, genomeScope2.0, BWA, GATK and SnpEff. These tools were involved in performing the different steps of the NGS analyses which included quality control check, genome size heterozygosity, alignment or mapping, variant calling and annotation. Nextflow was employed in this pipeline as a workflow management system and Docker was used for containerising all the tools and their software dependencies. The developed pipeline was then tested to verify its utility in NGS data analysis. Pipeline development is very important in genomics research because, it could help improve the quality and reliability of research outcomes and facilitate the sharing and comparison of data across different studies and research groups. Having a pipeline that can effectively be used in quick and simple analysis of genomes will significantly help in uncovering biologically meaningful or clinically significant variants. It is expected that the outcome of this study will significantly impact studies into the genetic basis of human diseases and precision medicine.

Keywords: *Next-generation Sequencing, Genomic analysis, Variant calling, Nextflow, Docker, Pipeline*