

**DEVELOPMENT OF A COMPUTATIONAL PIPELINE FOR THE
IDENTIFICATION OF NON-CODING RNAs FROM NEXT
GENERATION SEQUENCING DATA**

**NDIFON, NAOMI SIJE-OKIM
(22PBF02395)**

B.Sc Biochemistry, Bowen University, Iwo, Osun State, Nigeria

AUGUST, 2024

**DEVELOPMENT OF A COMPUTATIONAL PIPELINE FOR THE
IDENTIFICATION OF NON-CODING RNAs FROM NEXT
GENERATION SEQUENCING DATA**

BY

**NDIFON, NAOMI SIJE-OKIM
(22PBF02395)**

B.Sc Biochemistry, Bowen University, Iwo, Osun State, Nigeria

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

AUGUST, 2024

ACCEPTANCE

This is to attest that this dissertation is 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 declare that **I, NDIFON, NAOMI SIJE-OKIM (22PBF02395)**, conducted this research titled **“DEVELOPMENT OF A COMPUTATIONAL PIPELINE FOR THE IDENTIFICATION OF NON-CODING RNAs FROM NEXT GENERATION SEQUENCING DATA”**. It was carried out under the supervision of Dr. Itunuoluwa Isewon. Concepts of this research project are the results of the research carried out by NDIFON, Naomi Sije-Okim. Other researchers' ideas from published literature have been duly referenced.

NDIFON, NAOMI SIJE-OKIM

Signature and Date

CERTIFICATION

This is to certify that this dissertation titled “**DEVELOPMENT OF A COMPUTATIONAL PIPELINE FOR THE IDENTIFICATION OF NON-CODING RNAs FROM NEXT GENERATION SEQUENCING DATA**” is an original research carried out by **NDIFON, NAOMI SIJE-OKIM (22PBF02395)** in the Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota, Ogun State, Nigeria under the supervision of Dr. Itunuoluwa Isewon. We have examined and found this work acceptable as part of the requirements for the award of Master of Science (M.Sc.) in Bioinformatics.

Dr. Itunuoluwa M. Isewon
(Supervisor)

Signature and Date

Prof. Olufunke O. Oladipupo
(Head of Department)

Signature and Date

Prof. Afolashade O. Kuyoro
(External Examiner)

Signature and Date

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

Signature and Date

DEDICATION

This project is dedicated to God Almighty for showing up and showing out extraordinarily.

ACKNOWLEDGEMENTS

My heartfelt gratitude goes first and foremost to God, for the guidance, inspiration and resilience He provided throughout the execution of this work. I am also deeply grateful to the Covenant Applied Informatics and Communication Africa Centre of Excellence (CApIC-ACE) and the World Bank for sponsoring my studentship and giving me the opportunity to embark on this research and Covenant University for providing a conducive environment to think and work creatively.

I would also like to thank the Head of the Department, Prof. Olufunke Oladipupo and the entire faculty and student body of the Department of Computer and Information Sciences for their invaluable support and eager input into this work at all stages. I specially appreciate my Supervisor, Dr. Itunuoluwa Isewon, who is also the Postgraduate Coordinator of the Department of Computer and Information Sciences for her encouragement and support towards the success of this project.

The success of this project has been backed tremendously by the prayers and constant love of my family. To my parents; Dr. Marcellina Ezeh, Capt. Godwin Ndifon Ndifon, Mr. Emeka Ezeh, and Mrs. Esther Ndifon, thank you for the mental, spiritual, emotional and financial support over the last two years. For every textbook shared, every word and prayer of encouragement, every inspiring word, every strategic advice, every hand of help and every hour you took out to brainstorm my work with me even when it was outside the scope of your knowledge. You are my greatest blessing. And my siblings; Esther Chukwuemeka Ezeh, Marvel Ndifon and Jeff-Teddy Ndifon, thank you for being a pillar of love and support.

To friends and colleagues who have stood as anchors in this journey, I am also deeply grateful. I begin by specially thanking Paul Owolabi, Oluwatoba Alagbe and Ifiok Sampson for being so kindly receptive of all my stupid questions throughout this process. To Rifin Ashuza, thank you for spending hours patiently troubleshooting with me even at the expense of your own plans. To my dear friends; Samantha Ndifon, Bunmi Oladipo, Faith Nnaji, Jemimah Campbell, Angel Michael and Damilare Akinyera for always listening, always asking, “how can I help?”, always making me laugh when all else fails and encouraging me to try another day. Your support brought me this far. Thank you. I also thank my teammates at MEWC particularly Grace, Vivian, Rai, Agnes

and Tandi for being so encouraging and gracious in making sure my project remained a priority. I also specially thank the bioinformatics subreddit community for answering every stupid question with such graciousness and a genuine desire to help, and educate. And to my friends and colleagues at the Centre and the Department; Emmanuel Oba, Jumoke Ibitoye, Temilade Omonigbehin, Promise Onyemeachi, Temilayo Ladi-Lawal, Mercy Akinwale, Uche Nnaji, Ogooluwa Ogunpola, Julius Odunuga, Ademolu Ajao, Ifedayo Ajibola, Faith Adegoke, Blessing Onyido, Omokhose Judith Ojebuovboh, thank you for making this journey worthwhile.

And finally, and perhaps most importantly, I would like to thank myself; for getting up every morning without fail and walking worthy.

TABLE OF CONTENTS

CONTENTS	PAGES
TITLE PAGE	I
COVER PAGE	II
ACCEPTANCE	III
DECLARATION	IV
CERTIFICATION	V
DEDICATION	VI
ACKNOWLEDGEMENTS	VII
LIST OF FIGURES	XI
LIST OF TABLES	XII
LIST OF ABBREVIATIONS	XIII
ABSTRACT	XV
CHAPTER ONE: INTRODUCTION	1
1.1 Background to the Study	1
1.2 Statement of the Problem	4
1.3 Aim and Objectives	5
1.4 Research Methodology	5
1.5 Significance of the Study	6
1.6 Organization of the Dissertation	6
CHAPTER TWO: LITERATURE REVIEW	8
2.1 Preamble	8
2.2 Conceptual Review	8
2.2.1 Genomics and DNA Sequencing Techniques	8
2.2.2 Next Generation Sequencing	9
2.2.3 Non-Coding RNAs	10
2.2.4 Computational Tools for Identification and Annotation of ncRNAs	17
2.2.5 Scientific Workflow Management Systems	19
2.2.6 Containerization Tools	21
2.3 Review of Related Works	23
2.4 Summary of Findings	29
CHAPTER THREE: METHODOLOGY	30
3.1 Preamble	30

3.2	Materials	30
	3.2.1 Retrieval of Test Data	30
3.3	Methods	31
	3.3.1 Architecture of the Pipeline	31
	3.3.2 miRNA Identification	32
	3.3.3 lncRNA identification	35
	3.3.4 circRNA identification	37
3.4	Implementation of the Pipeline	39
CHAPTER FOUR: RESULTS AND DISCUSSION		42
4.1	Preamble	42
4.2	Results	42
	4.2.1 Containerization of Tools	42
	4.2.2 Development of the Pipeline	44
	4.2.3 Evaluation of the Effectiveness of the Pipeline	48
	4.2.4 Evaluation of the Performance of the Pipeline	48
4.3	Discussion	53
CHAPTER FIVE: CONCLUSION AND RECOMMENDATIONS		55
5.1	Conclusion	55
5.2	Limitations of the Study	55
5.3	Contribution to Knowledge	55
5.4	Recommendations	56
REFERENCES		57
APPENDIX		57

LIST OF FIGURES

FIGURES	TITLE OF FIGURES	PAGES
2.1	The History of DNA Sequencing	9
2.2	Next Generation Sequencing Techniques	10
2.3	Classification of non-coding RNAs	11
2.4	Biogenesis and Mechanism of Action of miRNAs	12
2.5	Structure and function of LncRNAs	13
2.6	Biogenesis of circRNAs	14
2.7	Synthesis of piRNAs	15
2.8	Biogenesis and Structure of snoRNAs	15
2.9	snoRNAs in Splicing of Primary Transcript	17
2.10	Characterization of Workflow Management Systems	21
2.11	Containerization Tools and their Features	23
3.1	Architecture of the Pipeline	32
3.2	Workflow of miRDeep2 module	34
3.3	Workflow of the PLEK tool	36
3.4	CIRI2 Algorithm	38
3.5	Workflow for execution of KNIFE	39
4.1	Snapshot of miRNA and lncRNA Identification Tools on Docker Hub	42
4.2	miRDeep2 on DockerHub	43
4.3	sRNAbench in sRNAtoolbox on Docker Hub	43
4.4	Successful Run of the Pipeline using Default Tools on Local Machine	43
4.5	Nextflow Generated Workflow Report for Pipeline	44
4.6	circRNA Identification Result	49
4.7	Distribution of Resource Usage in Pipeline	49
4.8	Distribution of Physical RAM Usage in Pipeline	49
4.9	Distribution of Virtual RAM Usage in Pipeline	50
4.10	Task Execution Time in Pipeline	51
4.11	CIRI execution runtime on local machine	51
4.12	PLEK execution runtime on local machine	51
4.13	miRDeep2 execution runtime on local machine	52
4.14	Input Read by Processes in Pipeline	52
4.15	Output written by Processes in Pipeline	53

LIST OF TABLES

TABLES	TITLE OF TABLES	PAGES
4.1	miRNA Identification Output Result by miRDeep2	45
4.2	Novel miRNAs predicted by miRDeep2	46
4.3	lncRNA Identification Output Result by PLEK	47

LIST OF ABBREVIATIONS

ASIR	Age-Standardized Incidence Rate
BC	Breast Cancer
BLAST	Basic Local Alignment Search Tool
BWA-MEM	Burrows Wheeler Aligner – Maximum Exact Match
BWT	Burrows-Wheeler Transform
CD-HIT	Cluster Database at High Identity with Tolerance
ceRNA	competing endogenous Ribonucleic Acid
CIGAR	Compact Idiosyncratic Gapped Alignment Report
circRNA	circular Ribonucleic Acid
CIRI	Circular RNA Identifier
CNV	Copy Number Variant
CPAT	Coding Potential Assessment Tool
CPC	Coding Potential Calculator
CPM	Counts Per Million
DAVID	Database for Annotation, Visualization and Integrated Discovery
DNA	Deoxyribonucleic Acid
FM-Index	Full text Index in Minute space
HPC	High Performance Computing
KEGG	Kyoto Encyclopedia of Genes and Genomes
KNIFE	Known and Novel IsoForm Explorer
LncDC	Long non-coding RNA Detection
LncRNA	Long non-coding Ribonucleic Acid
LSF	Load Sharing Facility
MGC	Maxim Gilbert Chemical Cleavage
miRNA	micro-Ribonucleic Acid
mRNA	messenger Ribonucleic Acid
ncRNA	non-coding Ribonucleic Acid
NGS	Next Generation Sequencing
ORF	Open Reading Frame
PEM	Paired End Mapping
piRNA	PIWI-Interacting Ribonucleic Acid

PLEK	Predictor of Long non-coding RNAs and mEssenger RNAs based on an improved k-mer scheme
RF	Random Forest
RFECV	Recursive Feature Elimination with Cross Validation
RNA	Ribonucleic Acid
rRNA	ribosomal Ribonucleic Acid
SGE	Sun Grid Engine
SLURM	Simple Linux Utility for Resource Management
SMRT	Single Molecule Real-Time sequencing
snNRNA	small nuclear Ribonucleic Acid
snoRNA	small nucleolar Ribonucleic Acid
SNP	Single Nucleotide Polymorphism
SVM	Support Vector Machine
TMM	Trimmed Mean of M-values
TNBC	Triple Negative Breast Cancer
TPM	Transcripts Per Million
tRNA	transfer Ribonucleic Acid
tsRNA	tRNA-derived small Ribonucleic Acid
UTR	Untranslated Regions
VM	Virtual Machine
WGS	Whole Genome Sequencing
XGBoost	Extreme Gradient Boosting

ABSTRACT

Recent advances in genomics have revealed the critical roles that non-coding RNAs play in disease occurrence, progression, and population disparities in patient treatment outcomes. With the evolution of Next Generation Sequencing (NGS) techniques and the generation of genomic big data, the ability of researchers to further explore the functions of these non-coding RNAs has become more widely accessible. However, efficient exploration requires user-friendly computational tools that can streamline and centralize data analysis, particularly for identifying non-coding RNAs within large volumes of NGS data. Current computational pipelines for non-coding RNA identification are often limited to detecting only a single class of non-coding RNA and do not integrate the latest standalone tools. Consequently, these pipelines are not workflow efficient as they restrict the comprehensive analysis of diverse non-coding RNA classes within a single framework. The aim of this study is to develop a computational pipeline for identifying multiple classes of non-coding RNAs namely micro RNAs, long non-coding RNAs and circular RNAs from NGS data. This aim was achieved by developing scripts for the selected software tools integrated into the pipeline and incorporating these scripts as individual processes within a unified Nextflow script. The software tools integrated into the pipeline include; miRDeep2, mirnovo and sRNAtoolbox for the identification of miRNAs; CIRC and KNIFE for the identification of circRNAs; PLEK and LncDC for the identification of lncRNAs. Nextflow was used as the scientific workflow management system and Docker was used for containerizing all the integrated tools and their software dependencies for easy use and reproducibility across different computing environments. The pipeline was then evaluated using test data provided by each of the individual software tools and it successfully identified all the reported miRNAs, lncRNAs and circRNAs, thus proving its effectiveness. Beyond the reduced execution time, the pipeline offers a more efficient solution by streamlining the analysis of non-coding RNAs and eliminating the need for separate software installation and environment setup, thereby reducing the user's workload.

Keywords: Next Generation Sequencing, Non-coding RNA, Nextflow, Docker, Computational Pipeline, micro RNAs, long non-coding RNAs, circular RNAs