DISSECTING MOLECULAR BIOLOGY THROUGH SINGLE MOLECULE SEQUENCING

NIRMAL SINGH MAHAR



DEPARTMENT OF BIOCHEMICAL ENGINEERING AND BIOTECHNOLOGY

INDIAN INSTITUTE OF TECHNOLOGY DELHI

JUNE 2026

© INDIAN INSTITUTE OF TECHNOLOGY DELHI (IITD), NEW DELHI, 2026

DISSECTING MOLECULAR BIOLOGY THROUGH SINGLE MOLECULE SEQUENCING

by

NIRMAL SINGH MAHAR

Department of Biochemical Engineering and Biotechnology

submitted

in partial fulfilment of the requirements of the degree of Doctor of Philosophy

to the



INDIAN INSTITUTE OF TECHNOLOGY DELHI

JUNE 2026

CERTIFICATE

This is to certify that the thesis entitled **'Dissecting molecular biology through single molecule sequencing'** being submitted by **Mr. Nirmal Singh Mahar** to the Indian Institute of Technology Delhi for the award of the degree of **'Doctor of Philosophy'**, is a record of the Bonafide research work carried out by him, which has been prepared under my supervision in conformity with the rules and regulations of the Indian Institute of Technology. The research reports and the results presented in this thesis have not been submitted for any degree or diploma in any other university or institute.



Dr. Ishaan Gupta
Associate Professor
Department of Biochemical Engineering and Biotechnology
Indian Institute of Technology (IIT) Delhi
Hauz Khas, New Delhi - 110016

Acknowledgements

Looking back to 2020, when I embarked on my PhD journey amid the COVID-19 pandemic, I knew it would rank among the most demanding yet fulfilling chapters of my life. This milestone would not have been possible without the invaluable support of key individuals and institutions over the past five and a half years.

Above all, I extend my deepest gratitude to my supervisor, Prof. Ishaan Gupta, for his unwavering guidance, encouragement, and expertise, which have profoundly shaped my research and broadened my understanding of the field.

I am equally thankful to my Students' Research Committee members, Prof. Shaikh Ziauddin Ahammad, Prof. D. Sundar, and Prof. Anita Roy, for their insightful feedback and steadfast support in refining this thesis.

My seniors played a pivotal role in my professional and personal development during my PhD years; I sincerely thank Dr Yogesh Kalakoti and Rohit Satyam for their mentorship, constructive criticism, and motivation that strengthened my work. I also appreciate the steadfast companionship of my lab mates and friends, Qazi Noorul Mateen, Anshul Budhraja, Dasari Abhilash, Juwayria, Khushi Chouhan, Pallavi Gupta, and Rajan, who provided constant encouragement.

A special note of profound thanks goes to Triveni Shelke, my anchor throughout this odyssey: my first reader, confidante, and unwavering partner, whose late-night reassurances reignited my purpose and affirmed my resolve.

Finally, my heartfelt appreciation goes to my parents for their boundless love, faith, and guidance, which fuelled my pursuit of excellence, and to my IIT Delhi PhD batchmates and friends who turned this chapter into cherished memories.



Nirmal Singh Mahar

Abstract

High-throughput short-read sequencing has transformed molecular biology, enabling applications from population-scale genomics to routine clinical diagnostics. However, its inherent read length limitation prevents accurate resolution of complex genomic regions such as repetitive elements and structural variants. Moreover, PCR amplification, integral to short-read workflows, erases native epigenetic modifications, necessitating separate, costly experiments to recover this information. Finally, because data analysis can only begin after sequencing completes, turnaround times are extended from hours to days.

Long-read sequencing addresses each of these limitations directly. It generates reads tens to hundreds of kilobases in length that span complex genomic regions, eliminates PCR amplification to preserve native base modifications, and streams data in real time, allowing parallel analysis of the generated data. Recent advances in Nanopore chemistry, coupled with the rapid evolution of AI-driven basecalling models, have elevated raw read accuracy to exceed 98%, steadily closing the gap with short-read platforms.

This thesis leverages contemporary Nanopore sequencing to resolve complex genomic architecture across three major subfields of genomics. The central objective is to demonstrate that Nanopore sequencing, alongside short-read sequencing, serves as a powerful genomic instrument capable of delivering highly contiguous, well-annotated assemblies and novel frameworks that harness its full capabilities.

It comprises three aims. 1) Using Nanopore sequencing at the whole-genome level to produce highly contiguous assemblies with better annotations and develop novel frameworks for uncovering complex regions of the genome; 2) Using Nanopore sequencing at the whole-metagenome level to produce highly contiguous metagenome-assembled genomes (MAGs) from environmental samples and develop a novel algorithm capable of capturing antimicrobial resistance genes (ARGs) directly from the sequencing data; 3) Develop a novel application utilising the real-time data generation advantage offered by Nanopore sequencing to capture the microbial composition by selectively targeting the universal phylogenetic marker.

Workflows, analysis and results presented in this thesis demonstrate the current standing of Nanopore sequencing and further emphasise its claim to resolve the complex regions of the genome. This thesis establishes Nanopore sequencing not as a niche tool for specialised applications but as a unified genomic instrument. The work conducted in this thesis, with improved validation, can be extended to clinical and environmental settings.

सार

उच्च-उत्पादन दर शॉर्ट-रीड अनुक्रमण ने आणविक जीवविज्ञान में क्रांतिकारी परिवर्तन किया है, जिससे जनसंख्या-स्तरीय जीनोमिक्स से लेकर नियमित नैदानिक परीक्षणों तक अनेक अनुप्रयोग संभव हुए हैं। हालांकि, इसकी अंतर्निहित रीड-लंबाई सीमा जटिल जीनोमिक क्षेत्रों, जैसे पुनरावृत्त तत्वों और संरचनात्मक वैरिएंट्स, का सटीक विश्लेषण करने में बाधा उत्पन्न करती है। इसके अतिरिक्त, शॉर्ट-रीड वर्कफ़्लो में अनिवार्य पीसीआर प्रवर्धन मूल एपिजेनेटिक संशोधनों को मिटा देता है, जिससे इस जानकारी को पुनः प्राप्त करने के लिए अलग और महंगे प्रयोग करने पड़ते हैं। अंततः, क्योंकि डेटा विश्लेषण अनुक्रमण पूर्ण होने के बाद ही आरंभ हो सकता है, परिणाम प्राप्त करने में कई घंटों से लेकर कई दिनों तक का अतिरिक्त समय लग जाता है।

लॉन्ग-रीड अनुक्रमण इन सभी सीमाओं का प्रत्यक्ष समाधान प्रस्तुत करता है। यह दसियों से लेकर सैकड़ों किलोबेस लंबाई तक की रीड उत्पन्न करता है, जो जटिल जीनोमिक क्षेत्रों को पूर्ण रूप से आच्छादित कर सकती हैं। यह पीसीआर प्रवर्धन की आवश्यकता को समाप्त करता है, जिससे मूल बेस संशोधन सुरक्षित रहते हैं, और यह वास्तविक समय में डेटा स्ट्रीम करता है, जिससे उत्पन्न डेटा का समानांतर विश्लेषण संभव हो पाता है। नैनोपोर रसायन में हालिया प्रगति तथा कृत्रिम बुद्धिमत्ता आधारित बेसकॉलिंग मॉडलों के तीव्र विकास ने असंसाधित रीड की सटीकता को 98% से अधिक तक पहुँचा दिया है, जिससे शॉर्ट-रीड प्लेटफॉर्म के साथ का अंतर लगातार कम हो रहा है।

यह शोध-प्रबंध जीनोमिक्स के तीन प्रमुख उप-क्षेत्रों में जटिल जीनोमिक संरचना को समझने के लिए समकालीन नैनोपोर अनुक्रमण का उपयोग करता है। इसका केंद्रीय उद्देश्य यह प्रदर्शित करना है कि नैनोपोर अनुक्रमण, लघु-रीड अनुक्रमण के साथ-साथ, एक शक्तिशाली जीनोमिक उपकरण के रूप में कार्य करता है, जो अत्यधिक सतत, सुव्याख्यायित असेंबली और इसकी पूर्ण क्षमताओं का उपयोग करने वाले नवीन ढाँचे प्रदान करने में सक्षम है।

इस शोधप्रबंध में तीन प्रमुख लक्ष्य समाहित हैं। पहला लक्ष्य सम्पूर्ण जीनोम स्तर पर नैनोपोर अनुक्रमण का उपयोग कर अत्यधिक सतत असेंबली तैयार करना, बेहतर एनोटेशन प्रदान करना तथा जीनोम के जटिल क्षेत्रों की पहचान के लिए नवीन ढाँचे विकसित करना है। दूसरा लक्ष्य सम्पूर्ण मेटाजीनोम स्तर पर नैनोपोर अनुक्रमण का उपयोग कर पर्यावरणीय नमूनों से अत्यधिक सतत मेटाजीनोम-असेंबल्ड जीनोम तैयार करना तथा अनुक्रमण डेटा से सीधे प्रतिजैविक प्रतिरोध जीन की पहचान करने में सक्षम एक नवीन एल्गोरिद्म विकसित करना है। तीसरा लक्ष्य नैनोपोर अनुक्रमण द्वारा उपलब्ध वास्तविक समय डेटा-

उत्पादन के लाभ का उपयोग करते हुए एक नवीन अनुप्रयोग विकसित करना है, जो सार्वभौमिक वंशविकासीय मार्कर को लक्षित कर सूक्ष्मजीव संरचना को शीघ्रता से पहचान सके।

इस शोधप्रबंध में प्रस्तुत वर्कफ़्लो, विश्लेषण और परिणाम नैनोपोर अनुक्रमण की वर्तमान स्थिति को दर्शाते हैं तथा जीनोम के जटिल क्षेत्रों को हल करने की इसकी क्षमता को और अधिक पुष्ट करते हैं। यह शोधप्रबंध नैनोपोर अनुक्रमण को किसी विशिष्ट अनुप्रयोग तक सीमित उपकरण के रूप में नहीं, बल्कि एक एकीकृत जीनोमिक साधन के रूप में स्थापित करता है। इस शोधप्रबंध में किया गया कार्य, उपयुक्त और सुदृढ़ प्रमाणीकरण के साथ, नैदानिक तथा पर्यावरणीय क्षेत्रों में भी विस्तारित किया जा सकता है।

Table of Contents

Certificate ..... **ii**

Acknowledgements ..... **iii**

Abstract ..... **iv**

सार ..... **v**

List of Figures ..... **x**

List of Tables ..... **xiii**

List of Equations ..... **xiv**

List of Abbreviations ..... **xv**

CHAPTER 1 Introduction and Thesis Overview **1**

1.1 Introduction to DNA Sequencing 1

1.2 Evolution of DNA sequencing technologies 2

1.2.1 First generation 2

1.2.2 Second generation 4

1.2.3 Third generation 6

1.3 Comparison of Long-read sequencing technologies 8

1.4 Technical Advancements in Nanopore Sequencing 10

1.5 Aims and Objectives 11

1.6 Thesis organisation 12

CHAPTER 2 Using Whole Genome Sequencing (WGS) to assemble and annotate high-quality genomes and complex genomic elements **14**

2.1 Introduction 14

2.2 Assembling and Annotating the Complete Genome of a novel *Pseudomonas entomophila* strain TVIN-A01 16

2.2.1 Introduction 16

2.2.2 Materials and methods 17

2.2.3 Results 19

2.2.4 Discussion 22

2.3 Assembling and Annotating Complete Genomes of *Candida auris* representative strains of three geographical clades 22

2.3.1 Introduction 22

2.3.2 Materials and methods 23

2.3.3 Results 25
2.3.4 Discussion 27
2.4 Genome-wide DNA N^6-methyladenosine in *Candida auris* reveals the potential role of epigenetic modifications in the acquisition of drug resistance 28
2.4.1 Introduction 28
2.4.2 Materials and methods 30
2.4.3 Results 32
2.4.4 Discussion 39
2.5 GenomeIndia: Assembling of 10,000 mitochondrial genomes from the Indian population 42
2.5.1 Introduction 42
2.5.2 Rationale and Objectives 44
2.5.3 A systematic comparison of human mitochondrial genome assembly tools 44
2.5.4 Nuclear-embedded mitochondrial (NuMT) DNA sequences landscape in the Indian population 63
2.5.5 ANOMALY: A Snakemake pipeline for identifying NuMTs from Long-Read Sequencing Data 69
2.6 Overall chapter conclusion 79
CHAPTER 3 Using Whole Metagenome Sequencing (WMS) to recover high-quality genomes with better genome annotations 82
3.1 Introduction 82
3.2 Assembling and annotating genomes recovered from an electroactive community enriched from wastewater from the Indian Institute of Technology Delhi, India 84
3.2.1 Introduction 84
3.2.2 Materials and methods 85
3.2.3 Results 86
3.2.4 Discussion 88
3.3 K-MARVEL: K-Mer-based Antimicrobial Resistance Virtual Exploration Lab, a novel algorithm to capture ARGs directly from raw sequencing data 91
3.3.1 Introduction 91
3.3.2 Materials and methods 92
3.3.3 Results 100
3.3.4 Discussion 110
3.4 Overall chapter conclusion 114
CHAPTER 4 Developing an open-source platform for the end-to-end analysis of 16S rRNA Nanopore sequencing data in real-time 116

4.1 Introduction 116
4.2 Materials and methods 120
4.2.1 Overview of NANOTAXI 120
4.2.2 Architectural design and core functionality 121
4.2.3 Synthetic community sequencing 123
4.3 Results 124
4.3.1 Real-time analysis 124
4.3.2 Cohort-level analysis 127
4.4 Discussion 132
4.5 Overall chapter conclusion 136
CHAPTER 5 Conclusion and future directions 138
5.1 Conclusion of the thesis 138
5.2 Future directions 140
5.2.1 Functional validation of *Pseudomonas entomophila* mutations 140
5.2.2 Mechanistic role of methylation in *Candida auris* drug resistance 140
5.2.3 Expanding the functional and population genetics of NuMTs 140
5.2.4 From MAGs to Microbial Community Engineering 141
5.2.5 Clinical Translation of K-MARVEL and NANOTAXI 141
CHAPTER 6 BIBLIOGRAPHY 142
CHAPTER 7 APPENDICES 176
7.1 Supplementary Methods 176
7.2 Supplementary Figures 183
7.3 Supplementary Tables 200
CHAPTER 8 DATA AVAILABILITY 202
CHAPTER 9 PEER-REVIEWED PUBLICATIONS 203
CHAPTER 10 RESUME OF THE AUTHOR 204

List of Figures

Figure 1-1: Evolution of sequencing technologies over two decades......2
Figure 1-2: Overview of the sequencing principles utilised by third-generation sequencing technologies......9
Figure 2-1: Schematic depicting the comparison of genome assembly outcomes using short-read versus long-read sequencing. 15
Figure 2-2: Key parameters for the assessment of the genome assembly......16
Figure 2-3: *Pseudomonas entomophila* pathogenesis and the characterisation of the virulence in the strain of interest......17
Figure 2-4: Phylogenetic tree based on the whole genome sequence that shows species cluster, subspecies cluster, per cent G+C, delta statistics, the genome size (in bp), protein count and ssu length (in bp).....20
Figure 2-5: IGV screenshots of missense mutations captured in the TVIN-A01 strain. 21
Figure 2-6: Phylogenetic Tree showing four distinct Geographical Clades of *Candida auris*......26
Figure 2-7: Comparison of key parameters to assess the quality of genomes assembled in the study and genomes obtained from the ATCC genome portal. 27
Figure 2-8: The motif sequence of 6mA sites detected by STREME for all four strains of *Candida auris*......33
Figure 2-9: Identification of 6mA sites in *Candida auris* Genomes......35
Figure 2-10: Distribution of 6mA sites in *Candida auris* Genes......37
Figure 2-11: Enriched Functional Categories Linked with Genes having high 6mA density. 39
Figure 2-12: GenomeIndia Project......43
Figure 2-13: Computation time depends on the number of threads and the size of the input data......54
Figure 2-14: Performance metrics of all the mitochondrial assemblers with simulated data as the input data. 55
Figure 2-15: Overview of the Results of the Qualitative Usability Evaluation. 56
Figure 2-16: Score of The Mitochondrial Assemblies Produced Using Simulated Datasets. 57
Figure 2-17: Comparing the Effect of Down Sampling on the Score of the Mitochondrial Assemblers......59

Figure 2-18: Comparison of the Effect of Down Sampling on the Perfect Assemblies Generated by the Assemblers. 59
Figure 2-19: Comparison of F1-scores based on SNPs called from assembled genomes and from sequencing data. 61
Figure 2-20: Process of numtogenesis and detection of NuMTs in GenomeIndia samples. 65
Figure 2-21: Distribution of NuMTs in the Indian Population. 67
Figure 2-22: Overestimation of complex NuMT length by short-read sequencing data. 68
Figure 2-23: ANOMALY workflow. 71
Figure 2-24: Complex NuMT event resolved by ANOMALY. 78
Figure 3-1: Comparison of short-read versus long-read sequencing in resolving complex metagenomes. 83
Figure 3-2: Stacked bar plot showing the relative abundance of species across two communities. 88
Figure 3-3: UpSet plot visualising the intersections of protein-coding genes across MAGs recovered from the two enrichment conditions. 89
Figure 3-4: Schematic of the K-MARVEL Algorithm. 97
Figure 3-5: Heatmap of classification performance metrics across different combinations of kmer and extension size used for running K-MARVEL across twelve datasets. 101
Figure 3-6: Comparison of the ARG classification performance between K-MARVEL and widely used read-based tools for both short-read and long-read datasets. 104
Figure 3-7: Comparison of the computational performance between K-MARVEL and read-based classification tools for both short and long-read datasets. 107
Figure 3-8: Computational Performance and Scalability of K-MARVEL. 110
Figure 3-9: Reconstruction of fragmented ARGs by K-MARVEL. 112
Figure 4-1: 16S rRNA gene as a universal phylogenetic marker. 118
Figure 4-2: Longer analysis time for processing short-read sequencing data. 119
Figure 4-3: Schematic of the backend analysis in NANOTAXI. 120
Figure 4-4: The real-time analysis interface of the NANOTAXI application. 125
Figure 4-5: Real-Time Analysis and Monitoring in NANOTAXI. 126
Figure 4-6: The cohort analysis interface of the NANOTAXI application. 127
Figure 4-7: Cohort-level analytical outputs generated by NANOTAXI 128

Figure 4-8: Volcano plots comparing each synthetic community against the ZymoBIOMICS reference community. 130
Figure 4-9: Functional inference and Enrichment analysis from PICRUSt2. 131

List of Tables

Table 1-1: Comparison between Oxford Nanopore sequencing and PacBio sequencing.....10

Table 2-1: SNPs found in the TVIN-A01 strain when compared to the reference L48 strain. 21

Table 2-2: Assembly Statistics and other information for *Candida auris* strains.....26

Table 2-3: Density of N^6-methyladenine (6mA) across the four strains of *Candida auris* representing its four geographical clades.....32

Table 2-4: Information about the mitochondrial genome assemblers utilised in this study. .47

Table 2-5: Scores of the assemblies generated by each mitochondrial genome assembler for samples with a mean sequencing depth of 10X. 60

Table 2-6: Performance comparison of NuMT detection tools on simulated datasets. 75

Table 2-7: Performance comparison of NuMT detection tools on real-world datasets. 76

Table 3-1: Assembly statistics of the metagenome-assembled genomes (MAGs) recovered from the electroactive community. 86

Table 3-2: Global Classification performance metrics of K-MARVEL for Long-read sequencing datasets with optimised k-mer and extension size (7, 12). 102

Table 3-3: Global Classification performance metrics of K-MARVEL for Short-read sequencing datasets with optimised k-mer and extension size (6,9). 103

Table 3-4: Global classification performance metrics of read-based ARG classification tools utilised in this study. 105

Table 4-1: Combination of classification methods and databases available to use in the NANOTAXI platform.....122

Table 4-2: Detailed information regarding the composition of the synthetic microbiome communities created for testing the real-time functionality of NANOTAXI. 124

List of Equations

Equation 2-1: Mitochondrial assembly score 51
Equation 2-2: Precision 52
Equation 2-3: Recall 52
Equation 2-4: F1-score 53
Equation 3-1: Composite score 95
Equation 3-2: Precision 98
Equation 3-3: Recall 98
Equation 3-4: F1-score 98
Equation 3-5: Accuracy 99
Equation 4-1: Good's coverage index 178

List of Abbreviations

API	Application Programming Interface
ARG	Antimicrobial Resistance Genes
ATCC	American Type Culture Collection
BH	Benjamini-Hochberg
BUSCO	Benchmarking Single-Copy Orthologs
CLR	Centred Log-Ratio
EC	Enzyme Commission
EET	Extracellular Electron Transfer
HAC	High Accuracy Model
KO	Kegg Orthologs
MAG	Metagenome-Assembled Genomes
NMDS	Non-metric Multidimensional Scaling
NuMT	Nuclear-embedded MiTochondrial segments
ONT	Oxford Nanopore Technologies
PacBio	Pacific Biosciences
PCA	Principal Component Analysis
PCoA	Principal Coordinate Analysis
PCR	Polymerase Chain Reaction
ROS	Reactive Oxygen Species
SMRT	Single-Molecule-Real-Time
SNP	Single Nucleotide Polymorphism
SV	Structural Variants
TSS	Total Sum Scaling
WGS	Whole Genome Sequencing
WMS	Whole Metagenome Sequencing