Application of Long-Read Sequencing Technologies to the Diagnosis of Pathogenic Fungal Species

Minh Thuy Vi Hoang

BMedSci (Hons I)

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy (Medicine and Health)

Faculty of Medicine and Health

The University of Sydney

2023

THE UNIVERSITY OF SYDNEY

Statement of Originality

This thesis is submitted to The University of Sydney in fulfilment of the requirements for the degree of Doctor of Philosophy (Medicine and Health).

I, Minh Thuy Vi Hoang, certify that to the best of my knowledge, the content of this thesis is original, except where otherwise indicated.

I certify that this thesis is the product of my own work and that the assistance received in preparing this thesis and sources have been acknowledged.

Parts of this thesis have been published or prepared for publishing in my name and authorship attributions for each manuscript are outlined in the preceding page. In addition to the statements above, in cases where I am not the corresponding author of a manuscript, permission to include the material has been granted by the corresponding author.

Signed,

Minh Thuy Vi Hoang

30th June 2023

Acknowledgements

At the completion of this momentous period of my life, I would first like to thank and acknowledge the many people that have provided me their invaluable guidance, help and support, without which this thesis would not have been possible.

I am deeply grateful to my primary supervisor Prof. Wieland Meyer for taking me on as an Honours student in 2018 and continuing to a Ph.D in 2019. Your passion and dedication to science have inspired me since our first meeting. Thank you for your endless encouragement, knowledge, and experience over the past few years which I will always look back on fondly.

I would like to express my sincere gratitude to my secondary supervisor Dr. Laszlo Irinyi for his guidance, patience and help throughout my academic research and daily life during my Ph.D. Your presence has always been a comfort inside and outside the lab which I will undoubtedly miss. Thank you for bleeding for our experiments and I wish the best for you in the future.

I would also like to offer my thanks to the members of the Medical Myology group. To Krystyna, my many thanks for providing me with hundreds of culture plates and chats over many cakes and glasses of wine. To Alex, thank you for your assistance with the database work, many trips to the hospital for SAB plates and your continued friendship. My best wishes go to all members of the group, and I hope one day we can all gather again.

Thank you to my collaborators Dr. Benjamin Schwessinger and Dr. YiHeng Hu for your hard work and for showing Laszlo and I around the beautiful ANU campus. Thank you to my coauthors Dr. Indira Basu, Prof. Sharon C.A. Chen, Assoc. Prof. Ariya Chindamporn, Dr. Marie Desnos-Ollivier, Dr. Dea Garcia-Hermoso, Abigail Graetz, Dr. Catriona Halliday, Prof. Paul D.N. Hebert, Dr. Menuk Jayawardena, Dr. Chun-Yi Lee, Prof. Celeste C. Linde, Dr.

Benjamin Mashford, Dr. Wendy McKinney, Ivan Mikhailovich Pchelin, Dr. Arthur J. Morris, Prof. Pietro Nenoff, Assoc. Prof. John P. Rathjen, Prof. Vincent Robert, Leon M. Smith, Prof. Tania C. Sorrell, Prof. Eric A. Stone, Eenjes Tavish, Dr. Chi-Ching Tsang, Dr. Silke Uhrlaß, Prof. Patrick C.Y. Woo, Assistant Prof. Chompoonek Yurayart and the ISHAM Barcoding of Pathogenic Fungi Working Group for the roles you have played in allowing this work to come together.

A huge thank you to the staff and students from the University of Sydney and the Westmead Institute for Medical Research. I would like to offer my special thanks to Dr. Cameron Webb, Naomi Hollier and Dr. Rogan Lee for reaching out to me and supporting me during this difficult time. Thank you to Ying Cao for her support and for booking many meeting rooms for me, to Dr. Aleksandra Petrovic Fabijam, Dr. Ali Khalid and Dr. Ali Bayoumi for their help in blood collection and to Joey Lai for his help with the tape station.

A special thank you to all the friends I have made during my time as a PhD student that have undoubtedly shaped who I have become as a person including Ariana, Christiane, Cristina, Darci, Diana, Dom, Erik, James, Helen, Laura, Kim, Moo Noi and Sam. A special thanks to Alma and Lana for the many long lunches, cups of tea, conversations and laughs which I will forever hold dear in my heart.

To my personal friends and family, thank you for your belief in me and support despite still not completely understanding what I do. To my parents, sister, and brothers, thank you for your sacrifice and support. I hope I have made you all proud. Lastly, to my partner James, you have been my absolute rock during this time, and have always been able to put a smile on my face. I thank you deeply for your endless encouragement, reassurance and understanding.

Words cannot describe how much it has meant to me to have everybody’s support and from the bottom of my heart, thank you.

Abstract

The advent of long-read sequencing technologies changed the landscape of next generation sequencing technologies which had previously been dominated by short-read sequencers. These can be applied using a metabarcoding approach, which involves the sequencing of amplified genomic regions, or metagenomic approach, in which all genomic material is sequenced. These new technologies, capable of ultra-long read lengths have a high potential for their application to fungal diagnostics, however, the application to the identification of pathogenic fungal species is currently limited. This thesis aims to contribute to the improvement of metabarcoding and metagenomic long-read sequencing of pathogenic fungal species.

Chapter 1 outlines the current routine diagnostic methods, the long-read sequencing technologies, and their application to fungal identification. Chapter 2 describes the generation of high-quality barcode sequences to improve the database of fungal reference barcodes. The ISHAM Barcoding database was extended in Chapter 3, to contain both primary and secondary fungal DNA barcode reference sequences. Chapter 4 compares the two barcodes and shows that the secondary fungal barcode improves identification in all analysed taxa, thereby emphasising its requirement in making DNA barcoding an essential tool for fungal identification. Chapter 5 establishes a new fungal-based bioinformatic tool using a machine learning approach designed for the identification of fungal species. Chapter 6 investigates a metagenomic sequencing approach using Oxford Nanopore Technology MinION flow cells to successfully identify fungal pathogens from clinical samples. The same approach but now using Flongle flow cells in Chapter 7 reveals the inability of this new flow cell to identify fungal pathogens in the clinical samples sequenced herein in its current state. Chapter 8 compares DNA extraction and human DNA depletion methods to develop an optimised

workflow that improves the identification of fungal spiked blood samples. In Chapter 9 bioinformatic tools and databases are compared to determine the use of fungal specific databases are most successful for fungal identification alongside the bioinformatic tool BLAST with optimised parameters. Chapter 10 summarises and contextualises the findings within this thesis and provides future directions to progress this research field.

List of Publications

Chapter 1

Hoang MTV, Irinyi L, Hu Y, Schwessinger B, Meyer W. Long-Reads-Based metagenomics in clinical diagnosis with a special focus on fungal infections. Front. Microbiol. 2021: 12:708550. doi: 10.3389/fmicb.2021.708550.

Chapter 3

Meyer W, Irinyi L, **Hoang MTV**, Robert V, Garcia-Hermoso D, Desnos-Ollivier M, *et al.* Database establishment for the secondary fungal DNA barcode translational elongation factor 1α (TEF1α). Genome. 2019;62(3):160-9. doi: 10.1139/gen-2018-0083.

Chapter 4

Hoang MTV, Irinyi L, Chen SCA, Sorrell TC, Group IBoMFW, Meyer W. Dual DNA barcoding for the molecular identification of the agents of invasive fungal infections. Front. Microbiol. 2019: 10:1647. doi: 10.3389/fmicb.2019.01647.

Chapter 5

Eenjes T, Hu Y, Irinyi L, **Hoang MTV**, Smith LM, Linde CC, Meyer W, Stone EA, Rathjen JP, Mashford B, Schwessinger B. Linked machine learning classifiers improve species classification of fungi when using error-prone long-reads on extended metabarcodes. [submitted to mBio].

Chapter 6

Irinyi L, Hu Y, **Hoang MTV**, Pasic L, Halliday C, Jayawardena M, *et al.* Long-read sequencing based clinical metagenomics for the detection and confirmation of Pneumocystis jirovecii directly from clinical specimens: A paradigm shift in mycological diagnostics. Med. Mycol. 2020;58(5):650-60. doi: 10.1093/mmy/myz109.

Chapter 9

Hu Y, Irinyi L, **Hoang MTV**, Eenjes T, Graetz A, Stone EA, *et al.* Inferring species compositions of complex fungal communities from long- and short-read sequence data. mBio. 2022: 13(2):1-16. doi: 10.1128/mbio.02444-21.

Appendix

Hoang MTV, Irinyi L, Meyer W. Long-read sequencing in fungal identification. Microbiology Australia. 2022;43(1):14-8. doi: 10.1071/MA22006.

List of Abbreviations

Abbreviation	Full Form
Al33F	EF1-1018F
Al33R	EF1-1620R
Al34F	EF1-1002F
Al34R	EF1-1688R
BAL	Bronchoalveolar Lavage
BDG	(1-3)-β-d-glucan
BLAST	Basic Local Alignment Search Tool
BOLD	Barcode of Life Data System
CFU	Colony Forming Unit
CNNs	Convoluted Neural Networks
CSF	Cerebrospinal Fluid
dsDNA	Double Stranded DNA
EtBr	Ethidium Bromide
HIV	Human Immunodeficiency Virus
IFD	Invasive Fungal Disease
INSDC	International Nucleotide Sequence Database Collection
IS	Induced Sputum
ITS, ITS1/2	Internal Transcribed Spacer Region
K2P	Kimura 2-Parametric Distance Model
LSU	Large Ribosomal Subunit
MALDI-TOF MS	Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry
mNGS	Metagenomic Next Generation Sequencing
NCBI	National Centre for Biotechnology Information
NCI	National Computation Infrastructure
NEB	New England Biolabs
NGS	Next Generation Sequencing
NH & MRC	National Health and Medical Research Council of Australia
nt	Nucleotide
ONT	Oxford Nanopore Technologies
OTU	Operational Taxonomic Unit
PacBio	Pacific Biosciences
PBS	Phosphate Buffered Saline
PCP	Pneumocystis Pneumonia
PCR	Polymerase Chain Reaction
PCR-RFLP	PCR-Restriction Fragment Length Polymorphism
PJP	*Pneumocystis jirovecii* Pneumonia
POI	Probability of Correct Identification
Q Score	Phred Quality Score
RAPD	Random Amplified Polymorphic DNA
ROC	Receiver Operator Curves

RPBI	*RNA Polymerase I*
RPBII	*RNA Polymerase II*
rRNA	16S Ribosomal RNA
rtPCR	Real Time PCR
S	Polymorphic Sites
SMRT	Single Molecule Real Time
SSU	Small Ribosomal Subunit
TEF1α	*Translational Elongation Factor 1α*
WIMP	What's In My Pot
WM	Westmead Medical Mycology Collection
ZMW	Zero-Mode Wareguiders
π	Nucleotide Diversity

List of Figures

Chapter 1

Figure 1: DNA Barcoding Gap

Figure 2: The primary fungal DNA barcode

Figure 3: The secondary fungal DNA barcode

Long-reads-based metagenomics in clinical diagnosis with a special focus on fungal infections

Figure 1: History of different identification techniques applied for the diagnosis of fungal pathogens

Figure 2: Nucleotide entropy plots showing the variations in the ITS1/2 regions for *Ascomycota*, *Basidiomycota* and *Zygomycota*

Figure 3: Proposed workflow for long-read-based sequencing in clinical diagnosis of fungal diseases

Figure 4: Tree of the Fungal Kingdom showing the major fungal classes and their human pathogenic species indicating species with (blue) and without (red) genomes

Figure 5: Potential cumulative increase in discriminatory power with the addition of more loci or whole genomes being used for identification

Chapter 2

Figure 1: Number of strains with fungal DNA barcodes generated

Figure 2: Current percentage distribution of strain representation of species in the ISHAM Barcoding Database divided by 1 strain, 2 strains and 3+ strains

Figure 3: Current frequency distribution of strains per species in the ISHAM Barcoding Database

Chapter 3

Database establishment for the secondary fungal DNA barcode translational elongation factor 1α (*TEF1α*)

Figure 1: Schematic structure of the primary (ITS) and secondary (*TEF1α*) fungal DNA barcode regions indicating universal primers for their application

Figure 2: Length distribution of secondary barcode (*TEF1α*) sequences in (A) GenBank and (B) ISHAM barcode database

Figure 3: Intraspecies variation for species, which are represented by more than three strains, in the ITS regions (white bars) – primary DNA barcode – compared with the *translational elongation factor 1α* (black bars) – secondary DNA barcode

Chapter 4

Dual DNA Barcoding for the Molecular Identification of the Agents of Invasive Fungal Infections

Figure 1: Current and potentially available techniques for fungal ID, showing the central role of the reference DNA barcode database and the pros and cons of all techniques

Figure 2: Schematic structure of the primary (ITS) and secondary (*TEF1α*) fungal DNA barcode regions indicating universal primers for their amplification

Figure 3: Intraspecies variation for species, which are represented by more than three strains, in the internal transcribed spacer (ITS) (blue bars) primary fungal DNA barcode compared with the *translational elongation factor 1α* (*TEF1α*) (grey bars) secondary fungal DNA barcode

Figure 4: Barcoding gap analyses for fungal taxa including causative agents of invasive fungal disease with intraspecies (grey) and interspeices (blue) genetic distances calculated using the Kimura 2-parameter model

Figure 5: Workflow of DNA barcoding based fungal ID (Steps 1 and 2 can be done in parallel to speed up the process)

Chapter 5

Linked machine learning classifiers improve species classification of fungi when using error-prone long-reads on extended metabarcodes

Figure 1: Visualisation of the fungal ribosomal DNA region and machine learning decision tree

Figure 2: Machine learning based species identification performs especially well for closely related species

Figure 3: Varying the confidence score threshold affects the recall rate and false positive rate when identifying target species from complex samples of unknown composition

Supporting Information Figure 1: The propagated confidence score is markedly different between spiked reads with known labels and reads of unknown origin

Chapter 6

Long-read sequencing based clinical metagenomics for the detection and confirmation of *Pneumocystis jirovecii* directly from clinical specimens: A paradigm shift in mycological diagnostics

Figure 1: The proportion of total sequence length for the 18 most common fungal genera identified by WIMP (A) and BLAST (B)

Figure 2: Comparison of fungal reads (A) and bacterial reads (B) classified by WIMP and BLAST from four induced sputum (IS) (C_1, C_2, C_3, Pj_1) and two bronchoalveolar lavage (BAL) specimens (Pj_2, Pj_3)

Figure 3: Schematic presentation of the nanopore based metagenomics pipelines to detect and characterize human pathogens from sample collection to data interpretation

Chapter 7

Figure 1: Summary of nanopore channels during sequencing of sample P2

Figure 2: The relative abundance of reads assigned to the most abundant taxon by WIMP classification for nanopore sequencing of patient samples P1(A), P2(B) and P3(C), excluding those assigned to *Homo sapiens*.

Chapter 8

Figure 1: The percentage of reads assigned to *Candida* after metagenomic nanopore sequencing of *C. albicans* spiked blood untreated and treated by human DNA saponin depletion

Figure 2: The percentage of *Candida* assigned reads after nanopore sequencing of untreated and saponin depletion treated *C. albicans* spiked blood at concentrations 10^7 CFU, 10^6 CFU, 10^5 CFU and 10^4 CFU

Chapter 9

Inferring Species Compositions of Complex Fungal Communities from Long- and Short-Read Sequence Data

Figure 1: Analysis of shotgun metagenomics data

Figure 2: Dynamics in precision, completeness, and remaining rate after applying progressive cutoffs on BLAST alignment metrics

Figure 3: Benchmarking of amplicon data sets

Figure 4: Improving community composition analysis by applying query coverage cutoffs

List of Tables

Chapter 1

Long-reads-based metagenomics in clinical diagnosis with a special focus on fungal infections

Table 1: Features of diagnostic tests available for fungal infections

Chapter 2

Table 1: Fungal species and the number of strains representing the species with the number of strains with ITS or *TEF1α* sequences generated

Chapter 4

Dual DNA Barcoding for the Molecular Identification of the Agents of Invasive Fungal Infections

Figure 1: Universal and fungal specific primers for the dual DNA barcoding scheme

Figure 2: Number of fungal strains with primary and secondary barcodes present in the ISHAM Barcoding Database analysed in this study

Chapter 5

Linked machine learning classifiers improve species classification of fungi when using error-prone long-reads on extended metabarcodes

Supporting Information Table 1: Sample labels, collection methods, DNA extraction methods and read counts before and after two-step data filtering

Chapter 6

Long-read sequencing based clinical metagenomics for the detection and confirmation of *Pneumocystis jirovecii* directly from clinical specimens: A paradigm shift in mycological diagnostics

Table 1: Sample characteristics, summary of nanopore sequencing and taxonomy assignments obtained by using WIMP and BLAST for four induced sputum (IS) (C_1, C_2, C_3, Pj_1) and two bronchoalveolar lavage (BAL) specimens (Pj_2, Pj_3)

Table 2: Number of reads classified at species level as fungi by WIMP and BLAST and their concordance in four induced sputum (IS) (C_1, C_2, C_3, Pj_1) and two bronchoalveolar lavage (BAL) specimens (Pj_2, Pj_3)

Chapter 7

Table 1: Quantification of extracted DNA purity from patient samples, using A260/A230 and A260/A280 values measured by spectrophotometry

Chapter 8

Table 1: The concentration and purity of DNA extracted from 10^8 CFU *C. albicans* spiked blood without human DNA depletion and with saponin depletion, buffer RLT depletion, plasma depletion, and Molzym depletion.

Table 2: DNA yield of *C. albicans* spiked blood untreated and treated by saponin depletion at concentrations 10^9 CFU to 10^4 CFU before nanopore sequencing library preparation and after library preparation, to be loaded for sequencing.

Table 3: The quality and quantity of extracted DNA from *C. albicans* spiked blood from concentrations 10^7 CFU to 10^4 CFU when untreated and treated using saponin human depletion protocols.

Table 4: The number of total classified reads, reads classified as *H. sapiens* and Reads classified as *Candida* spp. of metagenomic sequencing of extracted DNA from *C. albicans* spiked blood from concentrations 10^7 CFU to 10^4 CFU when untreated and treated using saponin human depletion protocols.

Chapter 9

Inferring Species Compositions of Complex Fungal Communities from Long- and Short-Read Sequence Data

Table 1: Metadata of the mock fungal community

Table 2: Characteristics of each sequence data set

Table 3: Assignment of published sequence data to genera after application of cutoffs to query coverage

Table of Contents

Coverpage _____ I
Statement of Originality _____ II
Acknowledgements _____ III
Abstract _____ V
List of Publications _____ VII
List of Abbreviations _____ IX
List of Figures _____ XI
List of Tables _____ XV
Table of Contents _____ XVII
Chapter 1 Introduction _____ 1
1.1 Mycoses _____ 2
1.1.1 Fungal Pathogens _____ 4
1.1.1.1 Cryptococcosis _____ 4
1.1.1.2 Pneumocystis _____ 5
1.1.1.3 Candidiasis _____ 6
1.1.1.4 Aspergillosis _____ 7
1.1.1.5 Mucormycosis _____ 7
1.1.2 Management of fungal diseases _____ 8
1.2 Traditional identification techniques _____ 9
1.2.1 Morphology-based methods _____ 10
1.2.2 Biochemical-based methods _____ 11
1.2.3 Serology-based methods _____ 12
1.2.4 MALDI-TOF MS _____ 12
1.3 Molecular Identification Techniques _____ 14
1.3.1 DNA Barcoding _____ 15
1.3.1.1 The Primary Fungal DNA Barcode _____ 17
1.3.1.2 The Secondary Fungal DNA Barcode _____ 19

1.4 Long-reads-based metagenomics in clinical diagnosis with a special focus on fungal infections 21
1.5 Hypothesis 42
1.6 Aims 42
1.7 References 43
Chapter 2 Reference DNA Barcode Sequence Generation 60
2.1 Introduction 61
2.1.1 GenBank 61
2.1.2 BOLD 62
2.1.3 RefSeq 63
2.1.4 UNITE 64
2.1.5 ISHAM-ITS Database 65
2.2 Methods 66
2.2.1 Fungal Culturing 66
2.2.2 DNA Extraction 66
2.2.3 ITS Barcode Amplification 67
2.2.4 *TEF1α* Barcode Amplification 68
2.2.5 Sequencing 69
2.2.6 Data Analysis 69
2.3 Results 70
2.3.1 Barcode Generation 70
2.3.2 Updated ISHAM fungal DNA barcode Database 74
2.4 Discussion 75
2.5 Conclusion 79
2.6 References 79
Chapter 3 Database Establishment for the Secondary Fungal DNA Barcode 83
3.1 Database establishment for the secondary fungal DNA barcode *translational elongation factor 1α* (*TEF1α*) 84
Chapter 4 Primary and Secondary Fungal DNA Barcode Comparison 95

4.1 Dual DNA Barcoding for the Molecular Identification of the Agents of Invasive Fungal Infections 96

Chapter 5 Bioinformatic Improvements in Fungal Metabarcoding 110

5.1 Linked machine learning classifiers improve species classification of fungi when using error-prone long-reads on extended metabarcodes 111

5.1.1 Abstract 113

5.1.2 Keywords 114

5.1.3 Background 114

5.1.4 Methods 117

5.1.4.1 Fungal pathogen sample collection, DNA extraction and ITS amplification 117

5.1.4.2 Library preparation and DNA sequencing using the MinION 118

5.1.4.3 Processing and manipulation of fungal pathogen reads 120

5.1.4.4 Augmenting read datasets 121

5.1.4.5 Generating consensus sequences for each species 121

5.1.4.6 Determining the relationships between samples 122

5.1.4.7 Creation of neural network classifiers to distinguish between samples 122

5.1.4.8 Evaluation of the machine learning classifiers 124

5.1.4.9 Chaining machine learning classifiers into a decision tree 124

5.1.4.10 Alternative methods for fungal pathogen read classification 124

5.1.4.11 Identifying a key species from a complex sample using machine learning 126

5.1.4.12 Identifying a key species from a complex sample using classifiers trained on simulated reads 127

5.1.5 Results 127

5.1.5.1 Design of a decision tree for machine learning classifiers for taxonomic assignment of fungal species 127

5.1.5.2 Machine learning methods provide robust classification, especially at the genus and species level 130

5.1.5.3 Identifying target species from a complex sample of unknown composition using the machine learning decision tree 133

5.1.5.4 Training a machine learning classifier with simulated reads resulted in similar accuracy to those trained on real data 135

5.1.6 Discussion 137

5.1.7 Conclusions 141
5.1.8 Declarations 141
5.1.9 References 143
Chapter 6 Nanopore Sequencing of Clinical Samples Using MinION Flow Cells 152
6.1 Long-read sequencing based clinical metagenomics for the detection and confirmation of *Pneumocystis jirovecii* directly from clinical specimens: A paradigm shift in mycological diagnostics 153
Chapter 7 Nanopore Sequencing of Clinical Samples using Flongle Flow Cells 165
7.1 Introduction 166
7.1.1 ONT Flongle 168
7.2 Methods 169
7.2.1 Sample acquisition and routine diagnosis 169
7.2.2 DNA Extraction for metagenomic sequencing 170
7.2.3 Library Preparation 172
7.2.4 DNA Sequencing 173
7.2.5 Data Analysis 173
7.3 Results 174
7.3.1 Routine Diagnosis of Patient Samples 174
7.3.2 Sample DNA Concentration and Purity 174
7.3.3 Metagenomic Sequencing 175
7.3.4 WIMP Classification 176
7.4 Discussion 179
7.4.1 DNA Extraction 179
7.4.2 Metagenomic Sequencing 181
7.4.3 WIMP Classification and Pathogen Detection 182
7.4.4 Study Limitations 184
7.5 Conclusions 184
7.6 References 185
Chapter 8 DNA Extraction and Host DNA Depletion Optimization 189
8.1 Introduction 190

8.1.1 Fungal Blood Infection Diagnostics 190
8.1.2 Metagenomic Sequencing and Host DNA Depletion 191
8.2 Methods 193
8.2.1 Comparison of Host DNA Depletion Methods 193
8.2.1.1 Blood Spiking 193
8.2.1.2 Host Depletion Methods 194
8.2.1.2.1 Untreated 194
8.2.1.2.2 Saponin Depletion 194
8.2.1.2.3 Buffer RLT Depletion 194
2.3.1.2.4 Plasma Depletion 195
8.2.1.2.5 Molzym Depletion 195
8.2.1.3 Manual Fungal Cell Wall Disruption 195
8.2.1.4 DNA Extraction 196
8.2.1.5 DNA Concentration and Purity 196
8.2.2 Limit of Detection of untreated vs saponin treated *C. albicans* spiked blood 197
8.2.2.1 Sample preparation 197
8.2.2.2 Nanopore Sequencing 197
8.2.2.3 Data Analysis 197
8.2.3 Optimisation of Saponin Human DNA Depletion 197
8.2.3.1 Blood Spiking 198
8.2.3.2 Saponin Host Depletion Method 198
8.2.3.2.1 Untreated 198
8.2.3.2.2 Saponin Depletion Method 1 198
8.2.3.2.3 Saponin Depletion Method 2 199
8.2.3.2.4 Saponin Depletion Method 3 199
8.2.3.3 Manual Fungal Cell Wall Disruption 199
8.2.3.4 DNA Extraction 200
8.2.3.5 DNA Purification 201
8.2.3.5.1 Zymo Research Genomic DNA Clean and Concentrator 201
8.2.3.5.2 AMPure Bead Cleanup 201
8.2.3.6 DNA Concentration and Purity 202
8.2.3.7 Nanopore Sequencing and Data Analysis 202

8.3 Results 202
8.3.1 Comparison of host depletion methods 202
8.3.2 Limit of Detection 203
8.2.3.1 Untreated 203
8.2.3.2 Saponin treated 205
8.2.3.3 Percentage of *Candida* reads 207
8.3.3 Optimization of saponin human DNA depletion 208
8.4 Discussion 211
8.4.1 Comparison of human DNA depletion methods 212
8.4.2 Limit of Detection 215
8.4.3 Optimization of Saponin Depletion 216
8.4.3.1 Saponin Depletion Method 1 217
8.4.3.2 Saponin Depletion Method 2 217
8.4.3.3 Saponin Depletion Method 3 218
8.4.4 Study Limitations 219
8.5 Conclusions 220
8.6 References 220
Chapter 9 Bioinformatic Improvements in Fungal Metagenomics - Inferring Species Compositions of Complex Fungal Communities from Long- and Short-Read Sequence Data 225
9.1 Inferring Species Compositions of Complex Fungal Communities from Long- and Short-Read Sequence Data 226
Chapter 10 Discussion 243
10.1 References 249
Appendix 251
Appendix 1 Strain List 252
Appendix 2: List of Equipment 280
Appendix 3: List of Computer Software 281
Appendix 4: List of Reagents 283
Appendix 5: Long-read sequencing in fungal identification 285

Chapter 1
Introduction

1.1 Mycoses

It is estimated that over 6 million fungal species are present on Earth [1, 2]. These species are ubiquitous throughout the environment and are able to survive in extreme conditions, such as underwater volcanoes [3], in subzero temperatures in the Arctic [4], the Chernobyl nuclear reactor [5], and in simulations of conditions on Mars in the international space station [6]. The huge diversity in the kingdom of fungi demonstrates the important role they play in the homeostasis of their environments through their vast interactions with bacteria, animals, and plants [7]. Fungi are also capable of having detrimental interactions with plants, animals, and humans, causing a large negative impact on society.

Approximately 8,000 fungal species are pathogens of plants that have a striking impact on agriculture, particularly through the infection of the five main food crops: rice, wheat, maize, potatoes, and soybeans. Statistics of low-level fungal diseases from 2009-2010 indicated that the loss of crops during 2011 could feed 8.5% of the world population [8]. *Magnaporthe oryzae* is an emerging pathogen capable of causing 100% wheat yield loss under the correct environmental conditions and is responsible for the destruction of 10-30% of rice crops annually [9, 10]. Wheat blast disease caused by *Magnaporthe oryzae* originated in Brazil. It has spread to Bangladesh in Southeast Asia, and rice blast disease is globally present, indicating the potential for global wheat and rice shortages [11-13]. *Botrytis cinerea* is another prolific plant pathogen, infecting over 200 diverse crop hosts, including vegetables, fruits, chickpeas, and legumes, causing 15-40% crop loss in harvested fruit [14, 15]. *B. cinerea* infections are controlled using various fungicides, however, widespread use has caused resistance to increase and spread [14]. Previous control strategies for agricultural fungal pathogens were centred on the application of fungicides. However, these have a negative impact on the environment, human and animal health, and increase resistance. That

is why there is a drive to replace fungicides with environmentally friendly antimicrobials and rigorous hygiene and quarantine practices.

Fungal species also have a detrimental effect on wildlife species, with 625 species attributed to the cause of infections [1]. Most notably, the fungal pathogen *Batrachochytrium dendrobatidis* has devastated the global amphibian population, threatening to cause the extinction of 50% of all amphibian species [16]. *B. dendrobatidis* infected amphibians have been found in the environment and in global trade, facilitating its spread worldwide [17, 18]. Control methods, such as increase in temperature and antifungal use have been demonstrated to reduce the incidence of *B. dendrobatidis* in small populations [19-21]. Further implementation of treatment methods and thorough surveillance will be necessary to control its global spread in the future [18]. Another prevalent fungal disease in wildlife is White-Nose Syndrome in bats, which emerged in 2006, caused by *Pseudogymnoascus destructans* [22]. This disease has led to a sharp decline in bat populations across North America, which will require decades to return to pre-white-nose syndrome population numbers [23]. Efforts to control the spread have been directed towards the treatment of the hosts and the reduction of the environmental pathogen reservoir. However, a widespread management plan has not been implemented [23]. Fungal infections in wildlife are difficult to manage, and comprehensive international efforts will be required to eliminate the pathogens, reduce the risk of host population extinction, and limit their global spread.

Fungal pathogens are estimated to cause 1.5 million deaths per year in humans, similar to tuberculosis and three times that of malaria [24]. The prevalence of mycoses has continued to increase over the past several decades. Despite these staggering numbers, the impact of fungal diseases on human health is often underappreciated [25]. Although different fungal pathogens have different risk factors, the general increase in incidence is attributed to a decrease in immunocompetence, leaving patients susceptible to infections they were

previously not vulnerable to [24]. A large driver of increased mycoses is the continuing HIV/AIDS pandemic, whilst underlying diseases are also common risk factors for fungal infection, such as diabetes mellitus, chronic obstructive pulmonary disease, tuberculosis, and cancers [26]. Additionally, aggressive treatment of diseases predisposes patients to fungal infections, as is the case for immunosuppressant use for whole organ and stem cell transplant patients and chemotherapy for cancer patients [27]. Treatment at medical centres is also a risk factor for fungal infection through exposure from invasive devices, such as catheters, inhalation of environmental spores and cutaneous contact with fungi found on surfaces [28].

1.1.1 Fungal Pathogens

The number of disease-causing fungal species is estimated to be over 600 species. However, the majority of severe infections are caused by a small number of species. Some of these are discussed below [1].

1.1.1.1 Cryptococcosis

Cryptococcosis is an important fungal disease predominantly caused by the members of the *Cryptococcus neoformans* and *Cryptococcus gattii* species complexes. There is an estimated one million new cases of cryptococcosis annually resulting in 600,000 deaths [27]. Infection is caused by the inhalation of fungal cells from an environmental source, which then generally infects the lungs and central nervous system. Infection may also occur anywhere else within the body. Over 80% of cryptococcosis is caused by the members of the *C. neoformans* species complex [29]. It is the most prominent fungal infection in HIV-infected people and has been classified as an AIDS-defining disease [30]. The environmental reservoir of *C. neoformans* has been identified as soil contaminated with bird guano and decaying wood. As such it is globally distributed [29]. Conversely the members of the *C. gattii* species

complex have environmental reservoirs in decaying would, most known are eucalyptus tree species and are found mainly in sub-tropical regions. Although *C. gattii* causes fewer infections than *C. neoformans*, it has been identified as an important pathogen, as it was found to be the cause of the cryptococcosis outbreak in healthy hosts in Canada and the USA, known as the Pacific Northwest outbreak [31]. Other *Cryptococcus* species, such as *Papiliotrema laurentii* (formerly: *Cryptococcus laurentii* and *Naganishia albida* (formerly: *Cryptococcus albidus*), have also been demonstrated to cause disease in immunocompromised hosts [32].

1.1.1.2 Pneumocystis

Pneumocystis jirovecii has emerged as an important pathogen causing *Pneumocystis jirovecii* pneumonia (PJP), accounting for an estimated 400,000 cases annually [24]. Originally thought to be a protozoan, *P. jirovecii* differs from other fungi by the presence of cholesterol in the cell wall instead of ergosterol [33, 34]. *Pneumocystis* species are host specific and have been identified in an array of mammals with *P. jirovecii* being the species causing infections in humans. Infection occurs through inhalation of *P. jirovecii* cysts and presents through non-specific symptoms, such as fever, shortness of breath, fatigue, and chest pain [34]. *P. jirovecii* spores are ubiquitous in nature and rarely cause disease in healthy hosts. However, PJP occurs in immunocompromised people. It was an initial concern during the early HIV/AIDS epidemic. However, the implementation of Highly Active Antiretroviral Therapy has lowered its burden in this population [35, 36]. More recently *P. jirovecii* infection is associated with immunodeficiency caused by autoimmune diseases, solid organ transplantation, stem cell transplantation, and cancer treatments [37]. The disease progression of PJP in non-HIV/AIDS patients is associated with acute onset, rapid progression, greater need for mechanical ventilation and intensive care, and as such poorer outcome [38, 39].

Prophylactic strategies have been studied and implemented to offset this increased risk [40]. As medical interventions continue to cause impaired immune systems in patients, the risk of PJP will continue to be a concern.

1.1.1.3 Candidiasis

Candida species are major fungal pathogens that cause nosocomial infections and fungal outbreaks [41]. Candidiasis presents in various ways, from common and less severe infections, such as skin infections, to severe invasive infections, amounting to approximately 400,000 cases of invasive candidiasis annually [24]. *Candida* species have been reported as the fourth most common cause of nosocomial and bloodstream infections and commonly cause hospital outbreaks [28, 42, 43]. A one-day international study of 14,414 patients in intensive care units found 18% of infections were caused by *Candida* spp. [44]. Additionally, the morbidity and mortality of *Candida* infections cause a large burden on society with an estimated 45-55% mortality rate in intensive care units and is estimated to cost $40,000 in United States Hospitals per patient [45, 46]. The majority (>90%) of invasive *Candida* infections are caused by *Candida albicans, Candida glabrata, Candida tropicalis, Candida parapsilosis* and *Pichia kudriavzevii* (formerly: *Candida krusei*) [43]. Although the global incidences vary with the different species, *Candida albicans* has historically been the most common cause of infections. However, the incidence of non-*albicans Candida* species has been on the rise in the past decade [47-50]. Species inherently resistant to common antifungals are increasing in incidence, such as *C. glabrata* and *Candida auris* [51, 52]. Different *Candida* species have specific risk factors, but general risk factors are central venous catheters, use of broad-spectrum antibacterial drugs, long-term intensive care unit stay, recent major surgery and immunosuppression [50].

1.1.1.4 Aspergillosis

Aspergillosis is a broad spectrum of fungal diseases from non-invasive to severe and fatal widespread dissemination caused by the members of the *Aspergillus* genus [53]. Although epidemiological information is limited, it is estimated that there are 200,000 invasive aspergillosis cases annually with high mortality rates of 30-95% [24]. Infection occurs after inhalation of fungal spores widely found in the environment. Although infection typically stems from the lungs, invasive aspergillosis may occur in any organ in the body. There are over 300 *Aspergillus* species currently described and the vast majority of invasive aspergillosis is cased by the species *Aspergillus fumigatus*, *A. flavus*, *A. terreus* and *A. niger* [53]. Most infections are caused by *A. fumigatus* as observed in a study of HIV-infected patients over 20 years in France [54]. Severe aspergillosis generally occurs in patients with compromised immune systems. However, it has also occurred in immunocompetent patients [53, 55-57]. In haematopoietic stem cell transplant patients, aspergillosis was the most common fungal infection seen and in solid organ transplant recipients an incidence of 10-15% was seen [58, 59]. Chronic obstructive pulmonary disease is also a prominent risk factor for aspergillus colonisation in the lungs alongside prolonged intensive care unit stays [60-62]. Treatment strategies have been established against invasive aspergillosis, however, there is a growing concern for increasing azole resistance being observed in *Aspergillus* species, especially *A. fumigatus* [63, 64].

1.1.1.5 Mucormycosis

Mucormycosis is a disease caused by fungi within the class Zygomycetes with more severe disease presentations being due to members of the order Mucorales [65]. Of these fungal species, the majority of serious mucormycosis infections are caused by the genera *Rhizopus* and *Mucor* [66]. This is a rare disease as there are only an estimated 10,000 mucormycosis

infections annually, however, it remains a concern as the mortality rates are ranging from 46-95% [66-68]. Infection typically occurs via inhalation of and/or inoculation of damaged skin or mucosa with fungal spores found throughout the environment [65]. As such, nosocomial infections occur due to increased spore load in the air caused by construction work, or introduction through medical procedures or contaminated instruments [69, 70]. Mucormycosis outbreaks may also occur as a consequence of natural disasters [71]. Infections manifest at different sites of the body, including the rhino-orbital-cerebral, cutaneous, and pulmonary regions, and advance to become disseminated [66]. Predisposing factors to mucormycosis are observed to vary amongst the developed and developing world with the developed world predominantly seeing infections in those with haematological malignancy, haematopoietic stem cell transplants, and solid organ transplants [72, 73]. Whereas in developing countries, particularly India, the prominent risk factor is uncontrolled diabetes mellitus [26, 74]. This has become a major concern during the COVID-19 pandemic, as there has been a 2-fold increase in mucormycosis infections in India from 2019 to 2020 [67]. India was already estimated to have a significantly higher rate of mucormycosis than developed countries, partially due to having the second largest population with diabetes mellitus, however, this has become exacerbated during the COVID-19 pandemic [75]. As such, mucormycosis is a prominent fungal disease with fast disease progression and high mortality rates.

1.1.2 Management of fungal diseases

The management of fungal diseases remains a global challenge. An effective strategy to limit the burden of fungal diseases would be the prevention of fungal infections. Vaccines are estimated to prevent six million deaths annually and their application against fungal pathogens would be greatly beneficial [76]. Currently vaccines in development target all

fungal species, as well as specific fungal genera, using different approaches, such as live attenuated, recombinant, and conjugate strategies [77, 78]. Despite these efforts, no fungal vaccines are currently in use or have been approved for use [77, 79]. To combat this the use of prophylactic antifungals has been implemented on potentially susceptible patients [80]. People undergoing medical treatment for underlying issues are those that would require prophylaxis. However, this poses a danger due to the potential negative drug-drug interactions [81-83]. A consequence of prophylactic treatment may be a breakthrough infection with a different fungal pathogen that was not targeted by the antifungal used, and so further work is required for the prevention of fungal diseases [84]. There are established treatment therapies for mycoses with the three main classes of antifungal agents being: polyenes, triazoles and echinocandins [85]. In recent years, resistance to these antifungals has been on the rise within fungal species and the frequency of infections in individuals with innate immunity has increased [86, 87]. As such, monitoring of antifungal use is required to prevent inappropriate use and resistance development. However, the practical implementation of such measures would be very difficult or impossible [85, 88]. To improve patient prognosis, rapid and accurate diagnosis is essential, as it enables faster implementation of the correct treatment [89, 90]. Fungal diseases often present with non-specific symptoms making diagnosis difficult, often leading to misdiagnoses.

1.2 Traditional identification techniques

Many identification techniques have been developed and are currently used in public health, animal welfare, plant protection, quarantine implications and many other industrial fields. These conventional diagnostic tools vary in their success, depending on the type of sample. Despite advancements in pathogen identification techniques, the traditional methods still have an important place in fungal identification.

1.2.1 Morphology-based methods

Morphology-based methods rely on the identification of phenotypic features of pathogens at macroscopic and microscopic levels. To visualize characterizing features, fungi are grown on solid and liquid media [91]. Growth at specific temperatures, on certain media compositions, and the length of incubation provides information for the identification of the causative pathogens. After the fungi have grown, they can be identified through their colour, size, shape, and growth pattern. However, these characteristics are often non-specific and can be shared amongst many species. As such, more in-depth observations are required, such as the complexity of morphology, the presence of sexual reproductive structures, the effect on surrounding media, and the effect of different staining techniques [92]. Additionally, fungi can be observed within patient samples prior to growth in separate media [93]. Tissue staining can select for pathogens and reveal identifying features, such as the effects on the host tissue, including: the location of infection, tissue inflammation and the presence of granulomas. Staining may also reveal features of the fungal invasion, such as the presence of cell walls or capsules, as well as pathogen cell clustering [94]. These are used as a preliminary identification that may be used to inform on treatment whilst cultures are grown, or other diagnostic techniques are employed.

The advantages of morphology-based diagnostic methods are these techniques can identify a large variety of previously well-described and characterized pathogens. The process is also simple and inexpensive when compared to more advanced techniques. However, pathogen growth can take up to several weeks for some fungal species, while rapid identification is essential in diagnostics [95]. These methods are demonstrated to be less sensitive than more advanced techniques as they are highly subjective and error-prone, leading to misidentifications [96, 97]. Accuracy of identification is largely based on the expertise of the

technician and so a high level of expertise is required [98]. Closely related species and species complexes are unable to be reliably identified as characteristic phenotypic traits are not displayed, which can only be genetically distinguished, e.g., within the genera *Fusarium*, *Aspergillus*, and *Scedosporium* [99, 100]. There are known species, such as the fungus *Pneumocystis jirovecii* that cannot be cultured under laboratory conditions and so morphological-based identification is necessary [101, 102].

1.2.2 Biochemical-based methods

Fungi can also be differentiated based on their biochemical profile [103, 104]. Currently, commercial biochemical-based identification instruments and tests are available for fungal identification. These tests provide positive and negative results for the presence of metabolic, enzymatic, and fermentation products produced by the fungal species, which are then used to provide an identification of the species. The available biochemical tests are widely varied in the number of fungal species able to be identified. Some platforms require the results to be read manually, such as the RapID Yeast Plus (Thermo Fisher Scientific, Waltham, MA, USA), whilst others are automatically read, such as the API ID 32C (bioMérieux, Marcy-l'Étoile, France) [94, 105, 106]. The VITEK 2 (bioMérieux, Marcy-l'Étoile, France) is a popular biochemical assay, that uses disposable cards, capable of identifying up to 50 common yeast species and characterise antifungal resistance profiles against common antifungals [107, 108]. The API 20C AUX (bioMérieux, Marcy-l'Étoile, France) is a widely used, automated biochemical assay that in a previous study from 290 clinical isolates correctly identified 83.7% yeast species, including 61,8% of rare yeast species [109].[111-113] Biochemical testing is limited in the range of fungal species it can identify and for those it has not been designed to identify, further testing is needed [109]. In a diagnostic setting this is not ideal as this increases the time for identification. Additionally,

the presence of biochemicals is often determined by a colour change and so the results are subjective and misidentifications can occur [110]. As such, biochemical testing is more suited towards confirmatory testing after a suspected pathogen has been identified.

1.2.3 Serology-based methods

Serology-based tests identify the presence of antigens and antibodies present within the patient's serum, indicating current and previous infections. The following serological tests are available: lateral flow assays, complement fixation tests, enzyme immunoassays, immunodiffusion assays and latex agglutination assays. These tests range in complexity, specificity, and accuracy. The complexity of serological tests also greatly varies. The commercial lateral flow assay, CrAg LFA (Immuno-Mycologics, Norman, OK, USA) is portable, equipment free and requires no refrigeration with a long shelf life [111]. Conversely, enzyme immunoassays are labour-intensive and require highly trained staff such as the commercial galactomannan test Platelia™ *Aspergillus* EIA (Bio-rad, Hercules, CA, USA). These methods are useful for diagnosis as they allow identification directly from patient samples and do not require separate pathogen growth [112]. Additionally, serological tests have positively identified pathogens where culturing gave a negative result [113]. However, serological diagnostic tests focus on the identification of one pathogen. The specificity and sensitivity of serological tests can be low, and a negative serological test does not exclude fungal infection. Serological testing is useful in fungal diagnostics as a screening tool in at-risk populations or in confirmatory testing where relevant clinical information is first considered.

1.2.4 MALDI-TOF MS

Matrix-Assisted Laser Desorption/Ionization time-of-flight mass spectrometry (MALDI-TOF MS) is a high throughput, reliable technology used to identify and analyse large biomolecules. Since its introduction in 2008, it has become a rapid and highly reliable analytical tool for the identification of diverse collections of microbes, as such it has been implemented in many diagnostic laboratories [114, 115]. MALDI-TOF MS requires microbial material for analysis which can be found directly in some patient samples with a high microbial load, such as urine or cerebrospinal fluid. However, this is commonly not available and so the growth of the fungus of pure cultures is required, which can encompass blood culture, liquid media, and pure cultures on solid media. A quick protein extraction is usually followed by purification to reduce factors that inhibit the ionization process or alter the result. This may be achieved through detergents, deionized water or commercial kits and reagents. The processed material is suspended in a crystalline matrix, vaporized, and ionized by a laser. A high voltage is then applied to excite the charged particles, and the time of flight of the particles is measured and recorded as the MALDI mass spectrum [116, 117].

MALDI-TOF MS has a low cost per sample (~$0.5) after the instrument (~$100,000) is acquired and has been demonstrated to be more cost-efficient than traditional identification techniques [118]. The identification based on generated MALDI mass spectra is more accurate than when using morphological and biochemical testing [119, 120]. The major drawback of MALDI-TOF MS is the heavy reliance on spectral databases for identification, with species not contained in databases given as unknown. Spectral databases are still lacking in coverage of pathogenic filamentous fungi due to difficulty in effectively lysing the tough cell walls using standard preparation techniques [120]. Commercial and laboratory methods are being developed to overcome this limitation and increase the representation of moulds in

spectral databases [115, 121]. Although suitable for major diagnostic laboratories, the high initial cost of the MALDI-TOF equipment prevents its widespread use.

1.3 Molecular Identification Techniques

The advancement in understanding the genetic information of microorganisms has been fundamental in the development of a range molecular diagnostic techniques. Polymerase chain reaction (PCR) involves double-stranded DNA being denatured into single strands and specific primers annealed, after which polymerase extends and amplifies the DNA. It revolutionised pathogen identification, as it enabled the amplification of DNA, which was previously not possible. It has now been used in combination with other identification techniques. Polymerase chain reaction – Restriction fragment length polymorphism (PCR-RFLP) involves the digestion of PCR amplicons with restriction enzymes, which are then visualised on an agarose gel to create distinct polymorphic fragments that are markers for identification. This method has been used for species-level identification [122-124]. Random amplified polymorphic DNA (RAPD) is another molecular technique where DNA is randomly amplified with short primers and the resulting bands are separated in an agarose gel from which a profile can be identified [125-127]. Nucleic acid probe hybridisation uses species-specific fluorescent or radioactive labelled DNA or RNA probes that are added to a sample to hybridise with complimentary strands. The resulting fluorescence or radioactivity is visualised [128, 129]. DNA fingerprinting involves DNA being cut by restriction enzymes, the resulting DNA fragments are then separated by length via gel electrophoresis and transferred to a membrane. The membrane is incubated with labelled probes targeting mini- or microsatellite sequences in the genome of an organism. The bands are then visualised by being exposed to x ray film and identified [130, 131]. Real-time PCR (rtPCR) differs from normal PCR by monitoring the amplification of DNA during the PCR process by measuring

fluorescence incorporated during the PCR process reflecting the quantity of the amplified DNA [132, 133]. These processes are capable of the identification of a causative organism. However, they must be optimised for each investigated fungal species. The T2Candida Panel (T2 Biosystems, Lexington, MA, USA) is a new diagnostic tool to identify invasive candidiasis from whole blood without culturing or nucleic acid extraction [134]. It involves PCR amplification and T2 Magnetic Resonance leading to the rapid, sensitive, and specific identification of candidemia [135, 136].

1.3.1 DNA Barcoding

DNA barcoding was proposed in 2003 by Hebert *et al.* and is the current gold-standard molecular identification technique for the identification of fungal pathogens [137]. It relies upon the extreme variation that occurs in DNA, an example being the potential 1 billion unique sequences made up of the four nucleotide bases within a 15 base pair length. DNA barcoding exploits this variation and relies upon the selection of short 500-800 base pair long DNA regions that discriminate between different species. An essential feature of DNA barcodes is the ability to generate barcoding gaps which describes the genetic distance between members of different species as being distinctly larger than the genetic distance between members of the same species [95] (Figure 1). The genetic distance between sequences is generated by calculating the pairwise sequence divergence using the Kimura 2-parametric distance model (K2P) [138]. DNA barcodes are generated by the amplification of the barcoding region using PCR, the resulting PCR product is then sequenced, generally using Sanger sequencing, and sequences are then identified by comparison to accurate species-specific reference DNA barcoding sequences. A characteristic of an ideal barcode would be the availability of robust universal primers alongside replicable PCR protocols, indicating the suitability of variable regions flanked by highly conserved flanking regions.

Simple alignment would also be desirable in a barcode to reduce complications when compared to reference barcodes.

Identification of fungal species using short DNA regions has previously been used in fungal taxonomy studies, featuring smaller datasets and a limited number of fungal species, indicating the suitability of DNA barcoding to the identification of pathogenic fungal [139-141].

DNA barcoding addresses the shortcomings of traditional identification techniques, as it relies on molecular methods that do not require prior pathogen growth, unlike culture-based methods which may require up to 4 weeks before adequate growth is achieved, leading to delayed treatment. The species resolution achieved by DNA barcoding is not possible with morphological techniques, as phenotypic characteristics do not always predict phylogenetic relatedness [142, 143]. They are highly laborious and require a high level of taxonomic expertise to provide accurate identification, a skill that is declining as there are fewer morphological taxonomy specialists.

Importantly, DNA barcoding generates data that lead to species hypothesis, as ideally novel species would be seen as diverging to separate species. Its application has led to the discovery of cryptic species and the decoding of species complexes. However, routine involvement of DNA barcoding in species delineation would require established thresholds for identification [144, 145]. DNA barcoding is highly suited toward the identification of pathogenic fungi within the clinical setting.

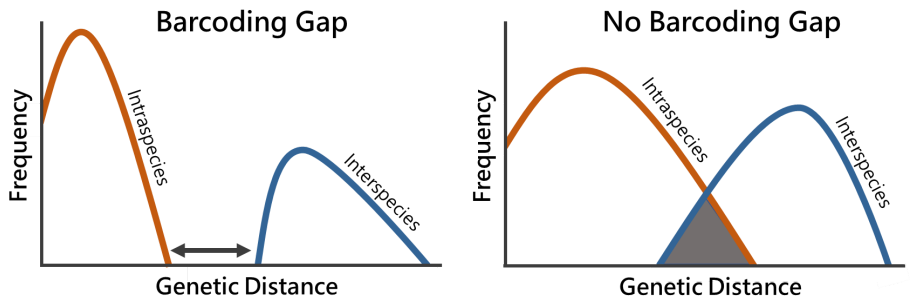


Figure 1: DNA barcoding gap. Visualisation of a barcoding gap where the intraspecies genetic distance is distinctly smaller than the interspecies genetic distance (left) and visualisation of an overlap between the intraspecies and interspecies genetic distances resulting in no barcoding gap (right).

1.3.1.1 The Primary Fungal DNA Barcode

The primary fungal DNA barcode has been established as the Internal Transcribed Spacer (ITS) region of the nuclear ribosomal DNA gene cluster [146]. The barcode consists of the central 5.8S gene, which is highly conserved, connected to the ITS1 region on the 3' end and the ITS2 region on the 5' end, which are the variable spacer regions [147]. These are flanked by the small ribosomal subunit (SSU) and the large ribosomal subunit (LSU), which are highly conserved (Figure 2). The ITS region was selected as the primary fungal DNA barcode after an international study involving multiple laboratories to evaluate six potential barcoding regions for the identification of all fungal species [146]. The six genomic regions analysed were the ITS, SSU, LSU, the largest subunit of *RNA polymerase I* (*RPBI*), the second largest subunit of *RNA polymerase II* (*RPBII*), and the mitochondrial maintenance protein *MCM7*. Over 700 fungal strains, representing 200 taxa, were used to determine the amplification success and probability of correct identification via these six regions.

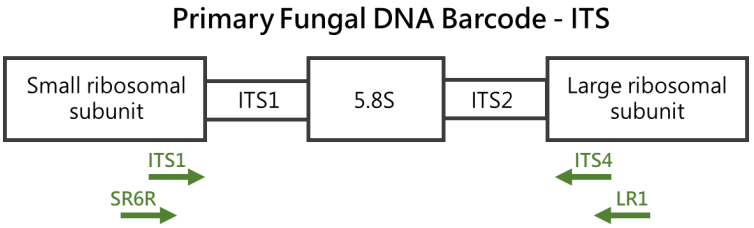


Figure 2: The primary fungal DNA barcode. The ITS region is the primary fungal DNA barcode and consists of the ITS1, 5.8S and ITS2 located between the small ribosomal subunit and large ribosomal subunit. The relative binding sites of common primer pairs, ITS1 and ITS4, and SR6R and LR1, are shown in green.

The ITS region reported the highest overall PCR success rate, and *RPB1* had the lowest. However, amplification success varied among fungal groups. ITS success varied from 65% in non-*Dikarya* lineages to 100% in *Saccharomycotina* and *Basidiomycota*, whilst *RPBI* varied from 14% in non-*Dikarya* lineages to 80% in *Saccharomycotina* of the phylum *Ascomycota*. *RPB1* had the highest overall probability of correct identification (POI) at 0.76 and had particularly high discrimination of *Ascomycota* and *Basidiomycota* which contain the majority of the currently described fungal species. ITS has a slightly lower POI with 0.73 and performed better than *RPB1* amongst Basidiomycota. In non-*Dikarya* lineages, LSU and SSU performed better than the ITS region. The choice of ITS as the primary fungal barcode was primarily based on its high amplification rate despite being surpassed by other regions in discriminatory power and although ITS had a lower POI of 0.73, this is comparable to the two-marker barcoding system used to identify plants, which has a POI of 0.70 [146, 148]. The ITS region is part of the nuclear ribosomal RNA gene cluster and as such has multiple tandem repeats within haploid genomes. As such, it requires less initial genetic material for amplification [143]. There are many ITS amplifying primer pairs available, such as the fungal-specific SR6R-LR1 pair and the non-specific ITS1-ITS4 pair, which limits the effect of any primer biases that may be present [147, 149]. Although many features of the ITS region lend itself to being suited toward its selection as the primary fungal DNA barcode, the

compromise on discriminatory power remains a concern as the ITS region is estimated to correctly identify 75% of medically relevant fungal species and is unable to generate clear barcoding gaps for some well-described species [95, 150, 151]. The intrastrain sequence heterogenicity amongst copies of ITS has been demonstrated to complicate chromatograph readings from Sanger sequencing, making a consensus sequence difficult to achieve [146]. Besides the many primers available for ITS it remains unable to identify approximately 10% of all fungal specimens tested [146, 150]. Strategies to alleviate the drawbacks of the single ITS fungal barcode were proposed to establish an alternate barcoding region or the addition of a secondary fungal DNA barcode to increase the coverage and accuracy of DNA barcoding for fungal species identification.

1.3.1.2 The Secondary Fungal DNA Barcode

Efforts to determine a new primary fungal DNA barcode to improve species identification largely concluded that different fungal groups require specialised identification efforts and so the alternative route of establishing a secondary fungal barcode was more realistic [95, 152-154]. A secondary barcode would ideally have high amplification success with universal primers and generate barcoding gaps in fungal taxa not possible to be identified with the ITS region. In such way, this additional barcode would supplement the primary fungal barcode. However, provided there was a large database of reference sequences, and it had adequate discriminatory power, the secondary barcode could also possibly function as a substitute for the primary barcode [151]. As such a large-scale study was undertaken by Stielow *et al.* resulting in the *translational elongation factor 1α* (*TEF1α*) being established as the secondary fungal DNA barcode [150]. The study analysed loci previously used in smaller taxonomic studies and new loci found using an *in silico* approach and a genomic search-based approach [155, 156]. Novel primer pairs were also established using similar methods for the new and

previously used loci. Potential barcodes were analysed with 1931 fungal strains, representing over 1500 fungal species, by amplifying the potential barcoding regions, bidirectionally sequencing the PCR products, and manual correction of sequencing artefacts, indels, and single nucleotide polymorphisms. The potential barcoding loci were then evaluated based on PCR success rate and the existence of barcoding gaps. Additionally, five loci were assessed using a mantel test of 3-dimensional plotting of pairwise distance matrices using 3 different fungal datasets. The five loci were also combined to generate a pairwise distance matrix, which was compared to the single matrices generated by a single locus. *TEF1α* demonstrated consistent amplification across different taxa and had an average success of 80% with poorer results for *Onygenales* and *Teratosphaeriaceae*. When looking at the most successful novel primer pair EF1-983F and EF1-1567R, the PCR success rate was 88%, which is comparable to the 92% success rate seen by ITS. *TEF1α* was also able to generate barcoding gaps in *Penicillium* and *Fusarium* species where the ITS region was unable to do so, indicating its higher discriminatory power for these fungal taxa. The 3-dimensional comparison of *TEF1α* had an optimum correlation with the overall matrix that was generated with all loci analysed. As such, *TEF1α* was selected as the secondary fungal DNA barcode as a compromise between discriminatory power and practical PCR amplification success [150]. *TEF1α* is a protein coding housekeeping gene and consists of regions of DNA separated by introns (Figure 2) [157]. It differs from the ITS region as there is only one copy in fungal cells and so amplification of protein coding genes is generally more difficult. However, the implementation of novel primers designed by Stielow *et al.* has mitigated this issue (Figure 2) [143]. It is a gene that slowly evolves and so effectively discriminates deep phylogenetic relationships whereas faster-evolving genes/genetic regions, such as the ITS region, discriminate more recent evolution events better, and as such the combination of ITS and *TEF1α* barcodes covers both [143]. The secondary barcode has also been used in previous

taxonomic studies within the Ascomycetes and that is why there are already reference sequences contained within public databases, allowing for immediate implantation, although a comprehensive quality-controlled collection of reference fungal barcode sequences would be more ideal [158-160]. The establishment and implementation of *TEF1α* is a forward step to increase the accuracy of DNA barcoding for the identification of pathogenic fungal species as an addition or alternative to the primary fungal DNA barcode ITS.

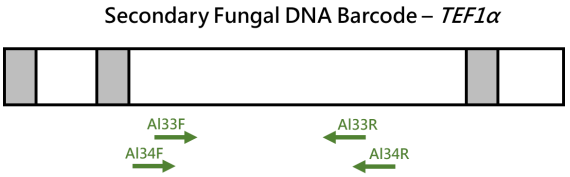


Figure 3: Secondary fungal DNA barcode. The *translational elongation factor 1α* (*TEF1α*) was selected as the secondary fungal DNA barcode to accompany the ITS region. It is composed of DNA regions separated by introns (grey) which may complicate identification. However, the primer pairs Al33F and Al33R, and Al34F and Al34R do not amplify these introns.

1.4 Long-reads-based metagenomics in clinical diagnosis with a special focus on fungal infections

The manuscript was published in Frontiers in Microbiology in January 2022 (Volume 12, article 708550, pages 1-42). Hoang M.T.V. is the primary author and wrote and corrected the manuscript. Irinyi L., Hu Y., Schwessinger B., and Meyer W. also wrote and corrected the manuscript. Irinyi L., Schwessinger B., and Meyer W. conceived the review. Meyer W. coordinated and supervised the review.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Minh Thuy Vi Hoang

30th June 2023

As a supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Professor Wieland Meyer

30th June 2023

frontiers in Microbiology

REVIEW
published: 06 January 2022
doi: 10.3389/fmicb.2021.708550

Long-Reads-Based Metagenomics in Clinical Diagnosis With a Special Focus on Fungal Infections

*Minh Thuy Vi Hoang[1,2], Laszlo Irinyi[1,2,3], Yiheng Hu[4], Benjamin Schwessinger[4] and Wieland Meyer[1,2,3,5]**

[1]Molecular Mycology Research Laboratory, Centre for Infectious Diseases and Microbiology, Faculty of Medicine and Health, Sydney Medical School, Westmead Clinical School, The University of Sydney, Sydney, NSW, Australia, [2]Westmead Institute for Medical Research, Westmead, NSW, Australia, [3]Sydney Infectious Disease Institute, The University of Sydney, Sydney, NSW, Australia, [4]Research School of Biology, Australia National University, Canberra, ACT, Australia, [5]Westmead Hospital (Research and Education Network), Westmead, NSW, Australia

OPEN ACCESS

Edited by:
Marcos Dias Pereira,
Federal University of Rio de Janeiro, Brazil

Reviewed by:
Leho Tedersoo,
Tartu University Hospital, Estonia
Sandip Paul,
JIS Institute of Advanced Studies and Research, India

***Correspondence:**
Wieland Meyer
wieland.meyer@sydney.edu.au;
wieland.meyer@curtin.edu.au

Specialty section:
This article was submitted to Microbial Physiology and Metabolism, a section of the journal Frontiers in Microbiology

Received: 12 May 2021
Accepted: 03 December 2021
Published: 06 January 2022

Citation:
Hoang MTV, Irinyi L, Hu Y, Schwessinger B and Meyer W (2022) Long-Reads-Based Metagenomics in Clinical Diagnosis With a Special Focus on Fungal Infections. Front. Microbiol. 12:708550. doi: 10.3389/fmicb.2021.708550

Identification of the causative infectious agent is essential in the management of infectious diseases, with the ideal diagnostic method being rapid, accurate, and informative, while remaining cost-effective. Traditional diagnostic techniques rely on culturing and cell propagation to isolate and identify the causative pathogen. These techniques are limited by the ability and the time required to grow or propagate an agent *in vitro* and the facts that identification based on morphological traits are non-specific, insensitive, and reliant on technical expertise. The evolution of next-generation sequencing has revolutionized genomic studies to generate more data at a cheaper cost. These are divided into short- and long-read sequencing technologies, depending on the length of reads generated during sequencing runs. Long-read sequencing also called third-generation sequencing emerged commercially through the instruments released by Pacific Biosciences and Oxford Nanopore Technologies, although relying on different sequencing chemistries, with the first one being more accurate both platforms can generate ultra-long sequence reads. Long-read sequencing is capable of entirely spanning previously established genomic identification regions or potentially small whole genomes, drastically improving the accuracy of the identification of pathogens directly from clinical samples. Long-read sequencing may also provide additional important clinical information, such as antimicrobial resistance profiles and epidemiological data from a single sequencing run. While initial applications of long-read sequencing in clinical diagnosis showed that it could be a promising diagnostic technique, it also has highlighted the need for further optimization. In this review, we show the potential long-read sequencing has in clinical diagnosis of fungal infections and discuss the pros and cons of its implementation.

Keywords: pathogenic fungi, long-read sequencing, metagenomics, identification, diagnosis, mycoses

INTRODUCTION

Rapid and accurate diagnosis of pathogens is essential in the management of infectious disease and is the ultimate goal for clinical microbiology laboratories. Successful diagnosis ensures effective treatment, with patient outcome improved the faster the pathogen is identified (Perfect, 2013). The ideal diagnostic test would be capable of reliably identifying any potential pathogen, provide additional information, such as the antimicrobial resistance profile, and offer the potential for further epidemiological analysis, while remaining cost-effective (Wickes and Wiederhold, 2018). The current available identification techniques for fungal pathogens are unable to meet all these criteria. The main drawback of the wide range of traditional identification techniques is that they rely in general on the *in vitro* growth of the causative pathogen before its diagnosis (Irinyi et al., 2016).

Sequence-based metagenomics is a powerful culture independent tool for the identification of mixed microbial communities, regardless of the ability of member organisms to be grown *in vitro* (Miller and Chiu, 2018). Recently, developed long-read sequencing technologies generate sequencing reads far longer than previous sequencing technologies and are capable of addressing the shortcomings of conventional identification, inaccurate, time consuming, requires specific expertise, and short-read based metagenomics, which is unable to detect all pathogens when based on short DNA fragments (Tedersoo et al., 2021).

This review explores previous and potential use of long-read sequencing in the clinical diagnosis, focusing on fungal infections.

NON-SEQUENCING-BASED IDENTIFICATION TECHNIQUES

Many identification techniques have been developed and are currently used in clinical diagnosis, public health, animal welfare, plant protection, quarantine implications, and many other industrial fields (**Figure 1**). These conventional diagnostic tools vary in their success depending on the type of sample. However, despite advancements in molecular identification techniques of pathogens, these traditional methods still are an important component of routine diagnosis.

Morphology

Morphology-based methods rely on the identification of phenotypic features of the pathogens at the macroscopic or microscopic level. Physical characteristics can be observed within patient samples through tissue staining. However, growth on media is often required for diagnosis (Guarner and Brandt, 2011; Austin, 2017). Fungal pathogens may be identified based upon simple characteristics, such as the color, size, and smell of the isolates. Often, these do not allow for complete diagnosis, as these features may be shared among many species. More in-depth observations are required, including the analysis of the complexity of morphology, the presence of sexual reproductive structures, the effect on surrounding media, and the effect of different staining techniques (Kurtzman et al., 2011; Schelenz et al., 2015).

Biochemical Methods

Commercial diagnostic tests based on characteristic biochemical profiles produced by fungi are available for diagnostic use. These tests provide a positive or negative result for metabolic, enzymatic, or fermentation products produced by specific fungal species and so are suitable for confirmatory testing after a suspected pathogen has been found through more broad range methods (Chen et al., 2014; Lackner et al., 2014; Schelenz et al., 2015; Wagner et al., 2017).

Serology

Serology-based tests identify the presence of antigens and antibodies within patient serum indicating current and previous infections (Im et al., 2019). Serological tests for fungal infections include testing for the presence of Galactomannan antigen for cases of invasive Aspergillosis (Pfeiffer et al., 2006; de Heer et al., 2019) and (1–3)-β-d-glucan (BDG) antigen testing for diagnosis of Candidemia and *Pneumocystis jirovecii* Pneumonia (Karageorgopoulos et al., 2011; White et al., 2017).

Targeted PCR Amplification

Targeted amplification of DNA regions using species-specific primers by PCR has been widely employed for the identification of fungal pathogens (Kidd et al., 2019). A wide array of primers targeting the ITS1/2 region of the rDNA gene cluster has been designed for specific fungal species, including the prominent pathogenic fungal species *Candida*, *Aspergillus*, and *Fusarium* (Carvalho et al., 2007; Springer and Loffler, 2017). Additional regions have also been targeted by species-specific primers, e.g., the D1/D2 region of the large ribosomal subunit (LSU) of the rDNA gene cluster is used for the identification of *Candida* species (Mannarelli and Kurtzman, 1998) and for *P. jirovecii* (Tia et al., 2012). The latest targeted amplification-based identification method is the T2Candida Panel (T2 Biosystems), which is a new diagnostic tool for rapid, sensitive, and specific detection of invasive candidiasis from whole blood without culturing or nucleic acid extraction (Neely et al., 2013). It relies on PCR amplification and T2 Magnetic Resonance which enables nanoparticle-mediated rapid detection of candidemia with a sensitivity of one colony forming unit (CFU)/ml of whole blood (Mylonakis et al., 2015; Clancy et al., 2018). The use of species-specific PCR is limited in fungal diagnostics, as it is reliant on prior suspicion of the fungal pathogen to select the correct PCR target, primers, and amplification conditions (Wickes and Wiederhold, 2018).

Matrix-Assisted Laser Desorption/Ionization Time of Flight Mass Spectrometry

Matrix-assisted laser desorption/ionization time of flight mass spectrometry (MALDI-TOF MS) is a high-throughput, reliable technology used to identify and analyze large biomolecules

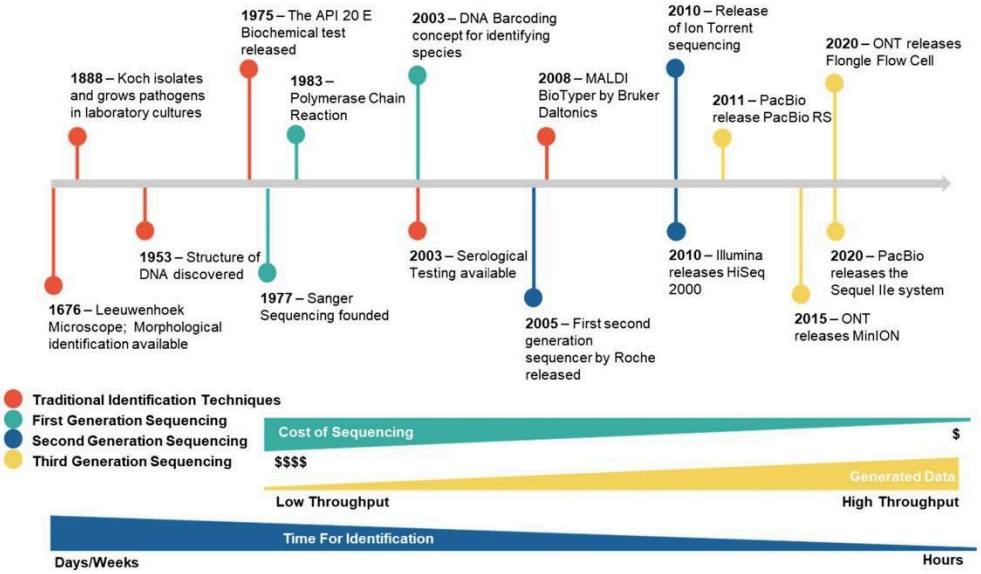


FIGURE 1 | History of different identification techniques applied for the diagnoses of fungal pathogens.

(Fenselau and Demirev, 2001; Clark et al., 2013). Starting material for MALDI-TOF can be found within sterile patient samples with high microbial load; however, this is not commonly possible and so prior pathogen growth is required. The processed material is then suspended in a crystalline matrix, vaporized, and ionized by a laser. Charged particles are then separated in a high voltage field, and the time of flight of the particles is measured and recorded as mass spectrum. Organism identification is obtained *via* the comparison of the obtained mass spectra with a reference database, which is currently the limiting factor (Angeletti, 2017).

Traditional identification techniques largely rely on prior growth of pathogens before testing may commence. This puts a strain on diagnosis as some fungal species may require up to 4 weeks before discriminatory features can be observed (Irinyi et al., 2016). Closely related species and species complexes, which may have different responses to antifungals, are unable to be reliably identified as characteristic phenotypic traits are not displayed and can only be genetically distinguished, such as in the genera *Fusarium*, *Aspergillus*, and *Scedosporium* (Goldstein et al., 2015; De Hoog et al., 2019). Traditional identification methods have been shown to be more insensitive than more advanced techniques, as results are highly subjective and error prone, leading to potential misidentifications (**Table 1**; Irinyi et al., 2016; Gustafson et al., 2019). Additionally, a number of fungal species, such as *P. jirovecii*, cannot be cultured under laboratory conditions and so culture-based methods are inappropriate for the identification (White et al., 2017). MALDI-TOF most closely meets the criteria for an ideal diagnostic test; however, it remains reliant on culturing and still lacks an exhaustive curated database. Although MALDI-TOF has been demonstrated to be more cost-efficient than traditional culturing techniques, the initial instrument cost limits widespread use in smaller diagnostic laboratories (Patel et al., 2017).

Overall, non-sequencing-based identification techniques are inadequate as the gold standard diagnostic technique for fungal infections.

SEQUENCING-BASED IDENTIFICATION TECHNIQUES

The advancement in the understanding of the genetic information of microorganisms has been fundamental for the development of molecular diagnosis techniques of pathogenic fungi. These techniques include PCR – restriction fragment length polymorphism analysis (Dendis et al., 2003), random amplified polymorphic DNA analysis (Brandt et al., 1998; Mobasherizadeh et al., 2016), hybridization with genus/species specific DNA or RNA probes (Sandhu et al., 1995; Lindsley et al., 2001), PCR-fingerprinting (Lieckfeldt et al., 1993), species-specific PCR assays (Martin et al., 2000; Kulik et al., 2004), real-time PCR (Klingspor and Jalal, 2006), and increasingly DNA sequencing (Pryce et al., 2003; Romanelli et al., 2010). DNA sequencing introduced by Sanger et al. (1977) dominated the sequencing technology for 20 years, being recognized as first-generation sequencing. When applied to the identification of

TABLE 1 | Features of diagnostic tests available for fungal infections.

	Diagnostic test	Available since	Culture dependency	Need for single culture	Test turnaround time without culture time	Accuracy	Curated database needed	Expertise	Cost	Portability	Resolution power	Potential for virulence and drug resistance detection	Generated clinical information
Non-sequence-based	Morphology	1676	Yes	Yes	Long	Low	No	High	Low	No	Medium	No	Low
	Biochemical	1975	Yes	Yes	Long	Low	Yes	Low	Medium	No	Medium	No	Low
	Serological	2003	Yes	No	Short	Low	No	Low	Low	No	None	No	Low
	Targeted PCR amplification (T2Candida panel)	2013	No	No	Short	High	No	Low	Low	No	High	No	Low
Sequence-based	MALDI-TOF	2008	Yes	Yes	Short	High	Yes	Low	Low	No	High	No	Low
	PCR based assays	1977	Yes	No	Medium	High	No	Medium	Low	No	High	No	Medium
	Sanger sequencing/DNA barcoding	1977	Yes	Yes	Medium	High	Yes	Medium	Medium	No	High	No	Medium
	Whole genome sequencing	2005	Yes	Yes	Long	High	Yes	High	High	No	High	Yes	Very high
	Long-read metagenomics/ Metabarcoding	2015	No	No	Short	High	Yes	Medium	Low	Yes	High	Yes	Very high

microorganisms, it relies on the analysis of pathogen-specific genome regions for comparison to reference sequence databases.

DNA barcoding is one of the most promising and efficient sequencing-based methods enabling rapid identification of species and recognition of cryptic species. The concept was first proposed by Hebert et al. (2003). DNA barcodes are standardized, easily amplified, universal, short DNA sequences (500–800 bp), which show a high divergence at the species level. These allow for the rapid identification of an organism though comparison of the generated genetic barcode to a reference collection of DNA barcodes from well-identified species. Two barcoding regions have been established for use in the identification of pathogenic fungi. The primary fungal DNA barcoding region is the internal transcribed spacer region (ITS1/2; Schoch et al., 2012) and the secondary fungal DNA barcoding region is the *translational elongation factor 1α* (Stielow et al., 2015). DNA barcoding using Sanger sequencing has major inherent limitations, including: (1) the requirement of a single fungal organism in the sample, (2) high target amplicon yield to avoid biases and errors (Polz and Cavanaugh, 1998), and (3) intra-individual variability (heteroplasmy). Especially, the high intra-genomic diversity of the ITS1/2 region in fungi cannot be simultaneously detected from one sample using Sanger sequencing alone, as samples with multiple amplicons will appear as mixed peaks in the sequencing chromatograms (O'Donnell and Cigelnik, 1997). The high intra-genomic diversity was only able to be detected after cloning (Jumpponen, 2003; Simon and Weiss, 2008) and the application of Next-Generation Sequencing (NGS) technologies (Ganley and Kobayashi, 2007; Colabella et al., 2021).

Sequencing technologies have recently rapidly advanced. The early NGS technologies were classified as second-generation sequencing technologies and achieved increased sequencing throughput after the amplification of thousands of DNA templates and the simultaneous sequencing of the resulting DNA fragments, saving time, and decreasing costs (Heather and Chain, 2016; Ravi et al., 2018). The most prominent short-read sequencing technology is those produced by Illumina. While Sanger sequencing suffered from low throughput, it did produce long-reads (up to 1,000 bp) with high accuracy. Short-read NGS technologies focused on achieving high throughput while sacrificing read length (100–600 bp; van Dijk et al., 2014; Illumina, 2021). The high throughput of NGS allowed for in-depth sequencing, recovering more data and thus more species from complex samples (Shokralla et al., 2014).

The introduction of sequence-based identification methods improved the speed and accuracy of fungal diagnostics and the rapid advancements of sequencing technologies enables the identification of fungi in complex samples.

METABARCODING AND METAGENOMICS

Metabarcoding/metagenomic sequencing is a method in which nucleic acid (DNA or RNA) of all organisms in a sample is extracted and sequenced using NGS techniques. The generated

sequences are then used to identify organisms present in the sample. It was first mentioned in 1998 (Handelsman et al., 1998), when it was applied for the culture independent analysis of a complex and diverse ("meta") community of microorganisms.

There are two main approaches which can be used to characterize the microbiome of a specific sample: (i) **targeted amplicon sequencing (metabarcoding)** and (ii) **shotgun metagenomics**. While the amplicon sequencing is based on the amplification of well-characterized genetic regions, shotgun metagenomics attempts to sequence the entire genetic content in a sample. The detailed differences between the two approaches have been previously discussed elsewhere (Forbes et al., 2017).

Targeted amplicon sequencing (metabarcoding) enables the direct identification of multiple species simultaneously from any environmental or clinical sample without culturing. It has been extensively used in microbiome studies and its biggest advantage over shotgun sequencing is the low amount of microbial DNA needed for identification and targeted enrichment of the sequence of interest. It is based on the amplification of specific taxonomic genetic regions (barcodes), such as the 16S ribosomal RNA (rRNA) gene cluster for bacteria (Riesenfeld et al., 2004; Janda and Abbott, 2007) or the internal transcribed spacer region (ITS1/2; Nilsson et al., 2009; Schoch et al., 2012) for fungi, by universal degenerated primers in a PCR reaction. These barcodes are ubiquitous in varying copy numbers, can be easily amplified, and have two fundamental characteristics: high taxonomic coverage and high resolution (Schoch et al., 2012). This approach combines the concept of DNA barcoding (Hebert et al., 2003) and the application of high-throughput sequencing technology (metabarcoding). The generated barcodes are then computationally clustered by sequence similarity into operational taxonomic units (OTUs) and queried against a reference database, such as UNITE[1] (Koljalg et al., 2005), BOLD[2] (Ratnasingham and Hebert, 2007), RefSeq[3] (Schoch et al., 2014), and the ISHAM Barcoding Database[4] (Irinyi et al., 2015; Meyer et al., 2019). The targeted approach only sequences the amplified barcodes and not the whole genome. Due to the smaller size of the generated data, the computational analyses are less complex and faster. However, amplicon-based sequencing has PCR biases due to its inability to amplify all microorganisms across multiple taxa and it does not provide any additional characterization beyond taxonomic information.

Short-read sequencing has been widely used in metabarcoding studies; however, it only targets short barcodes, such the ITS1 or ITS2 region for fungi, as second-generation sequencer read lengths (500–600 bp for paired and reads) are unable to span the full length of the established barcodes, the entire ITS1/2 region (Edouard et al., 2018; Hamad et al., 2018; Tsang et al., 2021). In addition, short-reads are unable to resolve repeated sequences that are longer than the reads generated by short-read NGS and as such these may lead to misassembles and gaps (Salzberg and Yorke, 2005; Treangen and Salzberg, 2011; Teng et al., 2017). Furthermore, even the short-reads lend themselves to identifying single-nucleotide variants and short indels accurately, longer structural variations are more challenging to resolve, as the variation is different within the ITS1 or ITS2 regions (**Figure 2**), leading to different discriminatory powers (van Dijk et al., 2018). As a result, in some species, the ITS1 or the ITS2 alone cannot resolve all species, which will only be possible using the combined ITS1/2 region (Irinyi et al., 2016). Longer sequencing reads would eliminate these issues and so drove the development of long-read NGS, see below.

In contrast, **shotgun metagenomic** sequencing allows for a higher resolution, as it sequences most parts of the genomes of every organism present in the sample. This feature enables the technology not only to identify the organism, but also to characterize their extended profiles, such as antimicrobial resistance, genetic subtypes, and virulence. It involves extraction of nucleic acid (DNA or RNA) from primary specimens (e.g., soil samples, water samples, blood, feces, bronchoalveolar lavage (BAL), and sputum, tissue), fragmentation of the obtained DNA or RNA extracts, library preparation, in-depth sequencing, and subsequent data analysis. Shotgun metagenomics is significantly more expensive (depending on the sequencing depth) and computationally more extensive than amplicon sequencing. Another challenge is the overwhelming amount of host or background DNA compared to the microbial DNA. Many methods have been developed for enriching microbial DNA to combat this for both environmental and clinical samples, see below. Additionally, in cases of non-sterile patient samples or co-infections, genomic material from multiple microbial species may prove difficult to assign correctly to the particular species present (Bharti and Grimm, 2021).

The application of shotgun metagenomics to diagnostics eliminates the need for culturing as identification would be made directly from patient samples (e.g., tissue, blood, feces, BAL, and sputum). As previously discussed, culturing greatly limits the detection of causative pathogens, as not all pathogens can be grown in a laboratory setting and it also greatly delays the timing of accurate identification (Garrido-Cardenas and Manzano-Agugliaro, 2017). In addition, metagenomics eliminates any sequencing bias that may occur as a result of PCR amplification (Wang et al., 2015). Comparing to amplicon sequencing, shotgun metagenomics also allows for better estimation of the relative abundance of microbes in samples based on the number of reads generated (Laudadio et al., 2018). Most importantly, shotgun metagenomics data contain additional genomic information from the patient samples, potentially informing not only on pathogen identities, but also underlying host factors. Previously established antimicrobial markers may also be identified from metagenomic data and as such can be used to guide targeted treatment, leading to the reduction of inappropriate use of antimicrobials and a subsequent reduction in the development of antimicrobial resistance. Additionally, further genomic information can be generated, to be used for epidemiological studies and outbreak tracing *via* the utilization of multi-locus sequence typing data generated as part of the metagenomics dataset (Couto et al., 2018).

[1]https://unite.ut.ee/
[2]https://www.boldsystems.org/
[3]https://www.ncbi.nlm.nih.gov/refseq/
[4]https://its.mycologylab.org/

The quantity and quality of data generated from metagenomic studies is reliant upon the NGS technology used and as such the advantages and disadvantages must be considered. The characteristics of the main technologies used for the identification of fungal pathogens are summarized in ***Table 1****.*

LONG-READ SEQUENCING TECHNOLOGIES

Third-generation sequencing technologies are characterized by the generation of ultra-long-reads and offer a number of advantages over short-read sequencing. Long-read sequencing routinely generates reads over 10 kb (Pollard et al., 2018; Mantere et al., 2019). Currently, commercial long-read sequencing is supported by two companies, Pacific Biosciences and Oxford Nanopore Technologies.

Pacific Biosciences

Pacific Biosciences (PacBio) released the first commercial long-read sequencer in 2011 and continue to maintain and release improved sequencing instruments (Ardui et al., 2018; Tedersoo et al., 2021). The basis of Pacific Biosciences sequencers is known as single molecule real-time sequencing (SMRT), which takes place in single use SMRT Cells. These contain multiple Zero-mode waveguiders (ZMWs) that hold an immobilized polymerase. Double-stranded DNA (dsDNA) is prepared for sequencing by forming SMRT bells by adding hairpin adaptors to both ends of the dsDNA template (Eid et al., 2009). Each SMRT bell is diffused into one ZMW and nucleotides with identifying fluorescent labels are added to the ZMW alongside reagents required for PCR. From there, the immobilized polymerase binds to the hairpin adaptor of the SMRT bells and begins replication. As a nucleotide is added to the complementary strand, fluorescence at a characteristic wavelength is emitted and the sequence of fluorescence pulses is recorded into a movie. The wavelengths emitted are then converted into a nucleotide sequence called a continuous long-read. After the polymerase completes replication of one DNA strand, it continues to sequence the opposite adapter and second strand. As a result, it is possible to generate multiple passes of the same template depending on the lifetime of the polymerase. Upon sequencing completion and conversion to continuous long-reads, subreads are formed by identifying and cleaving the adaptor region sequences. Multiple computationally merged subreads of the same template DNA strand are called circular consensus sequences or HiFi reads, and the more subreads, the higher the accuracy, currently reaching up to Q30–50 for consensus HiFi (Rhoads and Au, 2015; Tedersoo et al., 2021).

PacBio currently have three sequencers on the market, the Sequel System, the Sequel II system, and the Sequel IIe system, released in October 2020 (Pacific Biosciences, 2020). The newer iterations Sequel II and Sequel IIe boast a new SMRT Cell, dubbed the SMRT Cell 8 M, containing 8 million ZMWs capable of running for 30 h and generating up to 4,000,000 reads with over 99% (Q20+) accuracy (Wenger et al., 2019; Pacific Biosciences, 2021a). The Sequel IIe improves upon the Sequel II by upgrading the software and data analysis function through data processing capability on the instrument and cloud facilitation leading to potentially 70% reduction in secondary analysis time and 90% reduction in file transfer and data storage (Pacific Biosciences, 2020).

SMRT sequencers initially reported low read accuracy of 85–87%, when at the same time, short-read sequencers from Illumina had read accuracy >99% (Chin et al., 2013; Ardui et al., 2018). Errors in sequencing through insertions and deletions were more common in SMRT sequencing. However, improvements in sequencing chemistry and polishing methods increased the read accuracy of SMRT sequencing to 99.8% (Ardui et al., 2018; Tedersoo et al., 2018; Wenger et al., 2019). This allowed its use in direct sequencing studies rather than only be used to scaffold short-read Illumina reads (Parker et al., 2017; Teng et al., 2017; De Maio et al., 2019). The increased throughput of SMRT sequencers has made adequate coverage with multiplexed samples achievable (Armanhi et al., 2016; Pacific Biosciences, 2021b). The current SMRT Cell 8 M supports the simultaneous sequencing of 348 samples when using the largest barcoding kit PacBio currently offers, and the sequencing depth can achieve reference quality for *de novo* assemblies for genomes up to 2GB of a single sample and full length 16S region sequences for 192 samples for strain level identification (Pacific Biosciences, 2021b).

Disadvantages of PacBio instruments are, that they require a high initial investment cost, limiting widespread use. The throughput of Illumina sequencers remains higher than PacBio, as PacBio sequencers are limited by the number of ZMWs available in the SMRT Cells. Not all ZMWs are guaranteed to successfully sequence, as there maybe more than one DNA fragment being present in a ZMW, and there maybe issues with the polymerase binding. The nature of SMRT bells results in a trade-off between read length and read quality as longer DNA inserts are sequenced less often and hence the generated consensus sequence is of lower quality (Tedersoo et al., 2018). Further, the library preparation for the Sequel II Instrument is estimated to be 1 day, far longer than needed for the Oxford Nanopore Technology library preparation kits (see below), which may not be suitable for time sensitive diagnostic applications.

Overall, Pacific Biosciences has successfully created sequencing instruments capable of generating high accuracy ultra-long-reads.

Oxford Nanopore Technologies

The idea of nanopore-based technologies originated from the Coulter counter and ion channels (Cornell et al., 1997). The single molecule sequencing using biological nanopores was first proposed in 1995 and published in 1998 (Church et al., 1998). Kasianowicz et al. (1996) than described the detection of ssDNA passing through an α-hemolysin nanopore (2.6 nm in diameter; Kasianowicz et al., 1996). Oxford Nanopore Technologies (ONT) officially launched the concept of MinION™ in 2012 as a new generation nucleic acid sequencing technology based on nanopores, free from florescence labels and amplification requirements. It became commercially available in 2015 (Eisenstein, 2012; Deamer

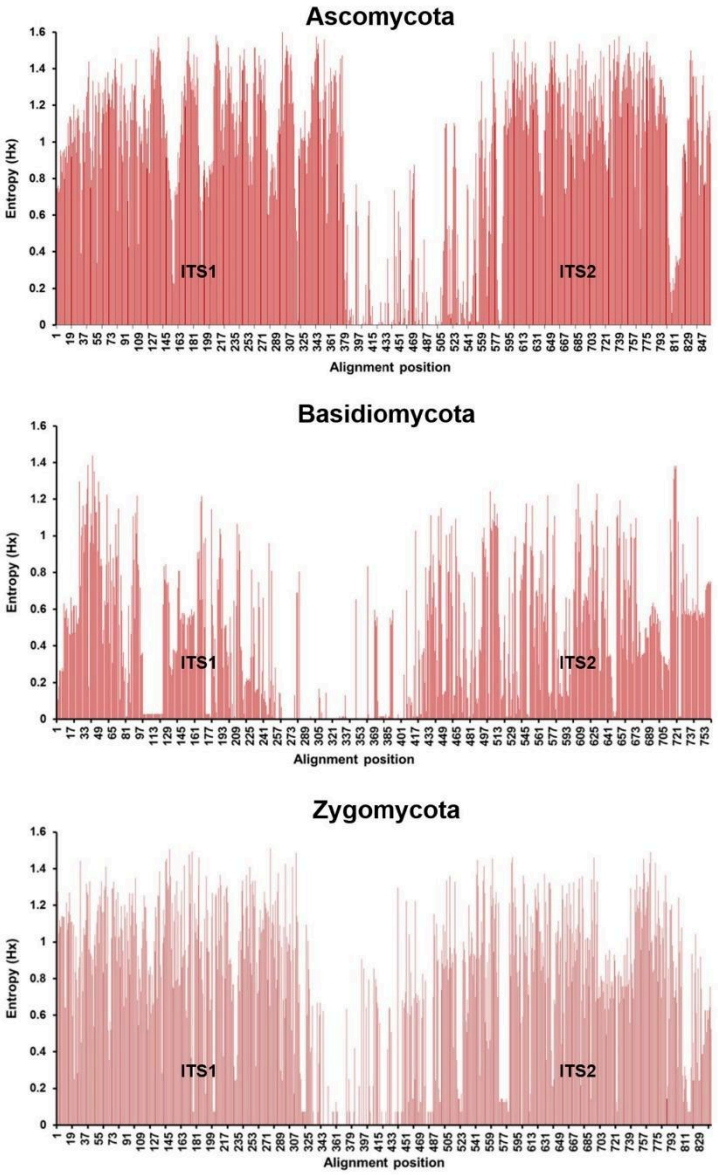


FIGURE 2 | Nucleotide entropy plots showing the variations in the ITS1/2 region for Ascomycota, Basidiomycota, and Zygomycota. The higher the bar is the greater the variation at that position in the nucleotide sequence. The variation of ITS1/2 is higher than that of the ITS1 or ITS2 regions alone and so indicates higher discriminatory power for identification. Entropy values were calculated using BioEdit 7.0.

et al., 2016). The MinION™ is a palm sized portable single molecule sequencing device capable of generating high-throughput, ultra-long sequence reads in real-time at relatively low cost.

The sequencing chemistry of ONT's sequencers has undergone rapid development with the latest flow cell released at the time of writing being the R10.4. However, the basis of nanopore

sequencing remains consistent, relying on biological nanopores embedded in solid state membranes within disposable flow cells (Leggett and Clark, 2017). Different sequencing library preparation methods are available depending on the intended use and data requirements. However, in all library preparation methods, a leader adapter and motor protein are ligated to the end of dsDNA fragments. This library is then loaded into the flow cells and the motor protein functions to unzip dsDNA and process the DNA strand throughout the pore at a specific speed. After one strand is processed, the pore is available to sequence the next available strand. Nucleotides are processed through the biological pores in groups of four called k-mers and as these travel through the pore, a change in an ionic current across the membrane occurs. These changes are then recorded as signal traces and during base calling, are converted to nucleotide sequences (Jain et al., 2016).

ONT's MinION flow cell contains 4x512 nanopore channels, each containing one nanopore, with 512 channels being accessed at a given time. These are run on both the MinION, which supports one flow cell at a time, and the GridION, which supports five flow cells. The PromethION is ONT's production-scale sequencer and provides 3,000 nanopore channels for each flow cell (Oxford Nanopore Technologies, 2021). All instruments can be run until the flow cells no longer have any pores available for sequencing and can be stopped when the intended sequencing depth is achieved. Additionally, the Flongle is a MinION adapter that enables the use of Flongle flow cells that contain 126 nanopore channels at a reduced cost (Oxford Nanopore Technologies, 2019).

The primary appeal of ONT sequencers over other current sequencing technologies is that read length is mostly dependent on sample DNA length, besides high DNA quantity and purity. It has been previously demonstrated to generate read lengths of over 2 million base pairs (Payne et al., 2019). ONT sequencers also offer reduced initial investment costs for sequencers when considering the MinION and the Flongle adapter although library preparation costs remain high (Tedersoo et al., 2021). Per sample sequencing costs depend on multiplexing, sequencing coverage, and re-use of flow cells after washing with a DNase. This makes sequencing available to a wider range of small routine laboratories when other sequencers may only be available for large laboratories or sequencing center. The availability of higher throughput instruments allows nanopore sequencing to be scaled to many different purposes. Multiplexing kits of up to 96 samples are available to further scale projects down if less in-depth sequencing is needed or a lower cost-per sample is necessary. ONT sequencers are highly suited for use in unconventional settings, as field library preparation kits are available that reduce library preparation time to 10 min with limited laboratory equipment. The sequencers are portable as the MinION and Flongle adapters are palm sized, allowing sequencing to be done outside the laboratory, such as at disease outbreak hotspots, war zones or for real-time environmental monitoring. This has previously been demonstrated through its use in space and resource poor settings (Quick et al., 2016; Castro-Wallace et al., 2017). The major drawback of nanopore sequencing has been the low read accuracy being reported with 90% with the R9.4 flow cell and earlier models (Lu et al., 2016). The introduction of new flow cells has led to strong improvements, with average accuracies of 87–98% being commonly reported (Loit et al., 2019; Logsdon et al., 2020). The highest accuracy ever reported after bioinformatics polishing, which might be computation-intensive and may take weeks, stands at >99% (Ashikawa et al., 2018; Morrison et al., 2020). In addition, new bioinformatic tools to improve the accuracy of nanopore reads are now readily available (Loman et al., 2015; Koren et al., 2017; Salmela et al., 2017; Xiao et al., 2017; Hu et al., 2021). As sequencing quality and length depends on the quality of the applied DNA/RNA samples a major drawback is that impurities within the loaded library may block pores and render them unable to sequence.

Overall, nanopore sequencing is by now a broadly accessible long-read sequencing platform with read length limited only by the DNA input making it a fundamentally different approach for sequencing.

HOW CAN LONG-READ SEQUENCING BE APPLIED TO CLINICAL DIAGNOSIS

As long-read sequencing platforms have the potential to accurately identify pathogenic species and simultaneously provide additional genomic information impacting patient outcome, we are focusing the remaining part of the review on the ONT and PacBio platforms for their use in clinical diagnostics.

The read length of SMRT sequencers has advanced to where it is comparable to those generated by ONT although read length is determined by input DNA for both sequencers. When using these directly on patient samples without prior amplification, as in metagenomics, long-reads provide genetic information spanning the whole informative regions of pathogen genomes. For example, DNA barcoding regions, which are between 500 and 800 bp, cannot be covered by short-read sequencing read lengths but are now being able to be fully sequenced by the long-read sequencing instruments. It would also be possible to sequence antimicrobial genes present in the genomes of resistant microbes, informing treatment strategies to improve patient outcome. Long-reads may also allow accurate sequencing of loci important in surveillance or strain typing. In samples where pathogen material is low in abundance, a prior amplification step of important genetic regions would ensure the sequencing of these important target regions although potentially missing other information.

The application of metagenomics to clinical diagnosis requires reliable and highly accurate sequencing and both ONT and PacBio have advertised over 99% per read accuracy for their latest sequencing chemistries. For PacBio, this has been reflected in many recent studies, e.g., Wenger et al., 2019. The accuracy of ONT depends on the type of sequencing chemistry and base-calling algorithms used. The accuracy of ONT technologies has been significantly improving and high accuracy levels are targeted to be achieved in the near future. Although nanopore sequencing may not yet be suitable for

applications, such as single-nucleotide variant identification or genotyping, which requires high coverage to exclude false positives, however, the accuracy required for pathogen identification can already be achieved (Greig et al., 2019; Feng et al., 2021).

Sequencing requirements of diagnostic laboratories cannot be guaranteed to be constant and so this may result in different throughput requirements. As such, investments into high-throughput sequencing instruments, such as the PacBio instruments, are not suitable for routine diagnostic laboratories. However, an alternative may be the implementation of core diagnostic centers that service multiple smaller laboratories and the multiplexing of samples, which would provide an effective way to reduce the cost-per sample. However, this would result in waiting for an adequate number of samples to be pooled for sequencing, which is contradicting the fact that fast diagnosis being the most important facture for the best patient outcome. As such, the smaller scale sequencing offered by ONT with the Flongle flow cells allows for sequencing of individual samples when needed, which is more suitable for general routine laboratories. Additionally, the initial capital investment in sequencing instruments, such as the ones produced by PacBio, may not be feasible for smaller diagnostic laboratories, which only could be overcome by pooling resources into larger diagnostic centers. The smaller scale ONT sequencers are generally affordable, making them more suitable in low throughput settings, such as it is the case in the diagnostic setting for mycoses.

The proposed workflow of the use of long-read sequencers in clinical diagnosis illustrates the need for timely identification (**Figure 3**). The time from sample collection to identification must be as rapid as possible and so the selection of a sequencer and its sample preparation and sequencing time must be considered. PacBio sequencing reports library preparation time of 1 day, while the slowest ONT library preparation kit, the direct cDNA Sequencing Kit, has an estimated preparation time of 270 min. The shortest library preparation option for ONT sequencing is the Field Sequencing Kit (SQK-LRK001) and the Rapid Sequencing Kit (SQK-RAD004), which have a preparation time of only 10 min although these kits produce a lower throughput (Kafetzopoulou et al., 2018). There are also further efforts to reduce turnaround time by reducing sample preparation and sequencing times (Loit et al., 2019). Further sample preparation may be added, which increases preparation times, such as a prior amplification step and sample multiplexing. Both PacBio (Sequel IIe) and ONT provide the option for real-time sequencing, thus eliminating the need to wait for the sequencing to be completed before base-calling can occur. This allows for a substantial reduction of the time from sample collection to pathogen identification to less than 24 h, enabling a faster introduction of the correct treatment, and improving patient outcome.

Before long-read sequencing can be applied to routine diagnostic testing, further work is required to determine the accuracy, specificity, and precision, of long-read sequencers in a clinical setting. The limit of detection must also be determined before it can be reliably implemented. In addition, adequate DNA isolation, library preparation, and loading according to standardized procedures performed by trained personnel are required to take full advantage of long-read technologies for routine diagnostics laboratories.

Long-read sequencers have many features that make them an attractive alternative for the use in clinical diagnosis of infectious disease agents.

LONG-READ SEQUENCING APPLIED TO INFECTIOUS DISEASE INVESTIGATIONS

The application of long-read sequencers to bacterial and viral species identification within a clinical setting is more advanced than that for fungal infections. The general smaller size of bacterial and viral genomes enables to use whole-genome

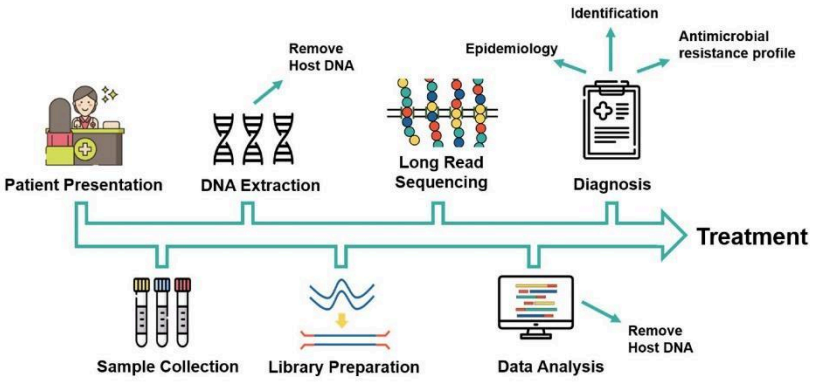


FIGURE 3 | Proposed workflow for long-read-based sequencing in clinical diagnosis of fungal diseases. Icons made by https://www.flaticon.com/authors/freepik.

sequencing studies instead of or in parallel with metagenomics studies, with both employing shotgun sequencing methods without prior amplification. As such, whole-genome sequencing studies will be discussed below alongside metagenomics and metabarcoding studies using long-read sequencers.

Bacterial Infections

The ability of PacBio to identify bacteria using the full length 16S region to the species level was demonstrated by the study by Earl et al. in which 100% of a mock community of 20 bacterial species was identified to the species level (Earl et al., 2018). This was then extended to apply sequencing of a mock community of 250 bacterial species, of which over 90% were accurately identified to the species level. Finally, it was successfully applied to characterize the microbiome of six sinonasal sites of 12 subjects (Earl et al., 2018). The use of PacBio in identification of bacterial species through sequencing of the full length 16S rRNA gene was further demonstrated through generation of microbiome profiles for vaginal samples and sputum samples from patients with cystic fibrosis (Hahn et al., 2016; Wagner et al., 2016). PacBio long-reads have also been used in the generation of whole-genome reference sequences of clinical samples indicating its potential use for future metagenomic studies to identify pathogenic bacteria (Kanamori et al., 2016; El-Rami et al., 2019).

Metagenomics and metabarcoding studies using nanopore sequencers for bacterial samples have been widely used in clinical and research settings. Rapid and accurate diagnosis of a *Capnocytophaga canimorsus* infection in a 62-year-old female patient using metagenomic nanopore sequencing has been reported (Bialasiewicz et al., 2019). This study demonstrated a positive identification was achieved by nanopore sequencing within 19 h, whereas conventional diagnostic techniques required 6.5 days to achieve the same. This short turnaround time from beginning of the sample preparation to pathogen identification is reflected in other studies, estimating a 6–8-h turnaround time (Schmidt et al., 2017; Charalampous et al., 2019; Gargis et al., 2019; Greig et al., 2019), and even using as little as 30 min of sequencing (Sakai et al., 2019; Taxt et al., 2020). Gu et al. (2021) also found a turnaround time of 6 h for nanopore sequencing compared to 24 h to Illumina sequencing. The study also demonstrated accurate identification of pathogenic bacterial species from a variety of clinical samples using metagenomic nanopore and Illumina sequencing. These samples also ranged from culture and PCR negative to PCR positive. In cases where NGS was unable to identify the causative pathogen, the cause was likely due to low pathogen titer or high levels of host background (Gu et al., 2021). An appeal of the use of metagenomics in clinical diagnostics is the potential for additional genomic information, such as antimicrobial resistance coding genes, to be obtained during sequencing. This has been achieved in studies for common bacterial pathogens, such as *Echerichia coli*, *Klebsiella pneumoniae*, *Mycobacterium tuberculosis,* and *Staphylococcus aureus* (Schmidt et al., 2017; Votintseva et al., 2017; Cheng et al., 2018; Charalampous et al., 2019; Gonzalez-Escalona et al., 2019; Taxt et al., 2020). Metagenomic sequencing of clinical bacterial samples has also revealed accurate epidemiological information that can be used for surveillance and outbreak tracing (Votintseva et al., 2017; Gorrie et al., 2018; Imai et al., 2020). ONT sequencers have also been used in metabarcoding studies involving the full length 16S rRNA gene to characterize complex mock communities and identify causative pathogens of clinical samples (Kai et al., 2019; Moon et al., 2019; Nakagawa et al., 2019; Winand et al., 2019).

> *Many published studies applying long-read sequencing to the identification of pathogenic bacterial species in clinical samples strongly suggest its potential incorporation into routine bacterial diagnosis and encouraging its potential application in fungal diagnoses.*

Viral Infections

PacBio sequencing has been used for whole-genome sequencing of a variety of human pathogenic viruses. A limitation of sequencing of viral genomes from clinical samples is the low viral load within cells. As such, reaching the required coverage to ensure an accurate consensus is a challenge. The small size of viral genomes lends themselves to overcome this issue through the amplification of the viral genome, which has been successfully used in whole-genome sequencing studies using PacBio, where the read lengths are able to span the whole amplicon length (Ocwieja et al., 2012; Dilernia et al., 2015; Bull et al., 2016; Nakano et al., 2017). PacBio sequencing has been demonstrated to be capable of the identification of viral pathogens from clinical samples but has not yet been used in routine viral diagnosis due to the high costs associated (Ho et al., 2017; Li et al., 2020). Identification of norovirus from patient samples has been performed to simulate diagnosis during an outbreak and demonstrated PacBio achieved sequencing of the entire viral pathogen, Norovirus GII, with high coverage of 99.11%; however, this was greatly limited by the long turnaround time (66 h; Li et al., 2020). PacBio reads have also been utilized in sequencing SARS-CoV-2 from patient samples to study the evolution and progression of disease and inform antiviral treatment (Ko et al., 2021). PacBio has additionally been used to achieve deep sequencing of genomic regions of viruses that are unable to be resolved using short-read sequencers (Bergfors et al., 2016; Huang et al., 2016; Ho et al., 2017; Lui et al., 2019; Takeda et al., 2019; Yamashita et al., 2020). Although these studies do not aim to identify the pathogens, the regions targeted for deep sequencing inform on resistance to antiviral treatments.

ONT has been widely used for the identification of viral pathogens in disease diagnostics and surveillance. Like the PacBio studies, nanopore sequencing of viral genomes often first involves either the amplification of the full genome or the generation of overlapping pooled amplicons that span the entirety of the genome (Quick et al., 2017; Deng et al., 2020). Nanopore sequencing has been demonstrated to be a useful tool for the identification of major pathogens during outbreaks, such as for the Ebola epidemic in West Africa from 2014 to 2016 and to generate additional genomic information on geographic, climate, and demographic factors (Hoenen et al., 2016; Quick et al., 2016; Dudas et al., 2017). Nanopore sequencing

has also been used in the recent COVID-19 pandemic, e.g., for a prospective study of the complex transmission within healthcare settings through the generation of 747 SARS-CoV-2 genomes from PCR-positive diagnostic samples within 24 h of sample collection (Meredith et al., 2020). Epidemiological information of the spread of a variant of SARS-CoV-2 was generated using nanopore sequencing to inform on further action needed to control its spread (Washington et al., 2021). ONT sequencers have additionally been used for whole-genome sequencing during the Zika virus epidemic (Faria et al., 2017), Yellow fever virus epidemic (Faria et al., 2018), and the Lassa fever outbreak in Nigeria (Kafetzopoulou et al., 2019). Whole-genome sequencing from clinical samples has also been demonstrated for human metapneumovirus (Xu et al., 2020), Chikungunya and Dengue viruses (Kafetzopoulou et al., 2018), Influenza (Lewandowski et al., 2019), and Hepatitis C virus (Greninger et al., 2015).

> *Long-read sequencing technologies have demonstrated their ability to be used to diagnose viral infections in the clinical setting and have a proven track record for analysis of samples during epidemics and a pandemic to produce valuable public health information.*

Fungal Infections

The application of long-read sequencing to fungal identification in a clinical setting is currently limited.

Whole genomes of major fungal pathogens have been generated with the use of long-read sequencers often in combination with short-read sequencing, so called as hybrid assemblies (Cuomo et al., 2017; Luo et al., 2017; Vale-Silva et al., 2017; Panthee et al., 2018; Rhodes et al., 2018; Morand et al., 2019; Schultzhaus et al., 2019; Pchelin et al., 2020). The use of whole-genome sequencing in clinical practice is not yet feasible, as the cost and data processing requirements are currently too high for routine use.

In clinical applications, PacBio sequencers have been used to characterize the fungal composition of fecal samples from 14 healthy individuals (Motooka et al., 2017). The study identified multiple fungal species present within the healthy gut mycobiota and characterized two major mycobiota types. This study by Motooka et al. (2017) additionally found PacBio was able to accurately identify all 26 fungal species in a mock community through metabarcoding of the ITS1 region, where short-read sequencers could not. Additionally, PacBio most accurately estimated the abundance of species within the mock community. The predominant reason PacBio outperformed the short-read sequencers was the use of long-reads, ensuring complete and accurate coverage of the entire 300–800 bp long ITS1 region (Motooka et al., 2017). PacBio has also been used for metabarcoding analysis of environmental fungal samples using the shorter ITS2 minibarcode and the large ribosomal subunit (28S), which are often used by short-read barcode studies (Cline and Zak, 2015; Kyaschenko et al., 2017). Sequencing of the full length ITS1/2 DNA barcoding region by PacBio sequencers has been demonstrated on a wide variety of fungal species (James et al., 2016; Walder et al., 2017; Heeger et al., 2018; Tedersoo et al., 2018; Purahong et al., 2019; Tedersoo and Anslan, 2019). A comparison between PacBio and nanopore sequencers found PacBio to be more efficient in the metabarcoding of complex environmental fungal samples due to errors found in nanopore sequencing (Loit et al., 2019). Additionally, sample to identification time was reduced to 2.5 h through metagenomic nanopore sequencing of plant samples (Loit et al., 2019). These studies indicate the further potential for long-read PacBio sequencing to be used in metabarcoding to identify causative pathogens in clinical samples.

The use of ONT in metabarcoding and metagenomics is similarly limited; however, direct clinical use has been previously demonstrated (Irinyi et al., 2020; Gu et al., 2021). Three patient samples, positive for *P. jirovecii* infection, alongside three samples without infection underwent metagenomic sequencing using ONT's MinION flow cells (Irinyi et al., 2020). The resulting reads were assigned using two different analysis tools, ONT's own "What's in my pot" and nucleotide BLAST, and the resulting assignments were compared. Sequencing of all three positive patient samples produced *P. jirovecii* reads while surprisingly, *Pneumocystis* reads were detected in negative samples when using the "What's in my pot" analysis tool; however, this is attributed to issues applying the analysis tool to fungal identification. *Homo sapiens* sequences accounted for 70–95% of all reads besides one outlier (10%). Additionally, the sample sites were bronchiolar lavage and induced sputum samples which are non-sterile sites and so multiple other fungal species were identified alongside the known causative pathogen. Notably, *Paracoccidioides* spp. were identified within the sample. However, this indicates a clear false positive, as *Paracoccidioides* spp. are geographically restricted to the tropical areas of Latin America, which did not align with the patient's travel history. This indicated that fungal diagnosis solely based off metagenomic sequencing using ONT has currently a number of limiting factors, including the low host:pathogen ratio, lack of comprehensive reference sequence databases, and the lack of appropriate bioinformatic tool, which need to be overcome before it can be applied in a routine diagnostic setting (Irinyi et al., 2020).

Gu et al. (2021) evaluated the diagnostic accuracy and performance of nanopore and Illumina sequencing on clinical body fluid samples (Gu et al., 2021). This study sequenced 87 patient samples by both sequencing technologies to compare the sensitivity and specificity. For fungal pathogen detection, the sensitivity and specificity of nanopore sequencing were found to be 90.9 and 100%, respectively, for nanopore sequencing, compared to 90.6 and 89.0% for Illumina sequencing. Additionally, five patient samples with negative culturing and PCR testing results were analyzed using NGS and three samples were found to be positive for fungal pathogens. For two samples, NGS did not detect any pathogens. However, the fungal pathogens *Cryptococcus neoformans* and *Sporothrix schenkii* were identified in these samples through further testing, most likely due to low pathogen titers or high host background DNA in the sample. This study demonstrated that mNGS can be applied to infectious disease diagnosis and would be a valuable tool in the routine diagnostic

process (Gu et al., 2021). Even though this study achieved the identification of causative pathogens using ONT sequencing, it also highlighted again many issues currently associated with practical implementation into routine diagnostics, including the importance of the quality of the input DNA, the proportion of pathogen DNA to host DNA, and the lack of comprehensive reference databases (Gu et al., 2021). Furthermore,

Metabarcoding for the identification of fungal pathogens from clinical samples using nanopore sequencing has also been performed. The ITS1/2 region was amplified in a patient sample previously confirmed having a *Candia albicans* infection and the resulting amplicons were sequenced on the ONT MinION (Ashikawa et al., 2018). This resulted in 97.3% of the first 4,000 reads and 99.1% of reads generated after 48 h being assigned to *C. albicans*, confirming the previous diagnosis. Another study used metabarcoding by nanopore sequencing to identify 16 members of an artificial fungal community using the ITS1/2 region (Mafune et al., 2020). The ITS1/2 region was also used to identify 87% of the fungal species present in 43 respiratory samples and identified others not detected through routine testing (Chan et al., 2020). Metabarcoding of 1,312 clinical respiratory samples with the bacterial 16S gene and fungal ITS1/2 region identified pathogens in 51.5% of samples with a turnaround time of <24 h. Nanopore sequence alone identified 31% of samples, while 0.5% were identified with culture-based techniques only (Wang et al., 2020). Long amplicon nanopore sequencing of the full ITS region and the whole fungal operon successfully identified all members of a mock community and microbial cultures (D'Andreano et al., 2021). The full ITS region was also used to characterize the fungal microbiota of healthy and potentially infected clinical canine samples (D'Andreano et al., 2021). The fungal intergenic spacer region has also been demonstrated to identify members from the *Cryptococcus gattii* and *Cryptococcus neoformans* species complexes by multiplexing 24 strains on a single MinION flow cell. Sequencing errors in regions with homopolymers were observed. However, with the R10.3 flow cell, this error rate was reduced by 57% and the sequence identity increased to 99.83% (Morrison et al., 2020).

> *These preliminary studies indicate the potential for the application of ONT and PacBio sequencers to the clinical diagnosis of fungal infections if the above-mentioned limitations are overcome to develop a workflow for clinical diagnosis.*

CURRENT ISSUES WITH LONG-READ SEQUENCING

Long-read sequencers have been demonstrated to effectively identify bacteria, viruses, and fungi, from clinical samples using both metagenomics and metabarcoding approaches. Although these studies have illustrated the potential for long-read sequencing in clinical diagnosis, a range of current challenges facing their implementation has also been highlighted.

DNA Extraction

Sequencing success depends on sample DNA quantity and quality. Pathogens are often present at low concentrations during infection and so high DNA concentrations are optimal to ensure adequate sequencing coverage of the causative pathogen for identification. Metagenomics and metabarcoding sequencing are aimed to represent everything within a sample and so ideally genomic material will be directly representative of what is present. As such, DNA extraction methods must be able to effectively lyse all cell types and preserve all genomic material.

For fungal pathogens, this presents a challenge, as fungal species have thick cell walls, containing polysaccharides and high content of lipids, that are difficult to lyse, which impedes the release of nucleic acids. As such fungal DNA extraction methods often involve a manual disruption step, such as bead bashing or crushing with liquid nitrogen to break open the fungal cell walls. This often leads to DNA shearing, greatly reducing the lengths of DNA fragments, defeating the purpose of long-read sequencing. A number of methods describing the extraction of DNA from fungal material for the use of nanopore and PacBio sequencing have reported to produce DNA fragments with an average fragment size above 35 kbp, although this has not yet been applied to clinical samples (Feehan et al., 2017; Schwessinger and Rathjen, 2017; Inglis et al., 2018). The development of better extraction methods is important to improve the application of long-read sequencers to accurate clinical diagnosis.

Ratio of Host to Pathogen DNA

A major problem in metagenomics analysis is the high amount of host (human) DNA in the sample, which may shadow the detection of pathogen DNA. One possible solution would be to increase the proportion of pathogen DNA in comparison with host DNA within the sample. This may be achieved through the depletion of human DNA in the clinical sample, which still is suboptimal even though there are several commercial kits and laboratory protocols for this purpose available (Marotz et al., 2018; Charalampous et al., 2019; Helmersen and Aamot, 2020; Heravi et al., 2020).

An alternative to this would be to increase the abundance of pathogen DNA, which can be achieved through amplification of important genomic regions, such as in metabarcoding studies, or amplification of overlapping DNA regions as done in viral whole-genome sequencing (Quick et al., 2017; Deng et al., 2020). An improved host to pathogen DNA ratio would greatly improve identification using long-read sequencing as it would enable higher coverage of the causative pathogen, leading to more accurate identification and the possibility of further genomic information. An increase in pathogen reads would also make long-read sequencing more cost-effective since more relevant data would be generated, possibly allowing for multiplexing of samples.

Another way to enrich pathogen DNA and deplete host DNA is the application of adaptive sequencing with nanopore sequencers. DNA molecules are selected for by identifying unwanted sequences and reversing the voltage across the membrane, thereby stopping

sequencing. Bioinformatic tools have been developed to implement adaptive sequencing for the selection of specific chromosomes and the depletion of bacterial genomes to enrich the remaining species in a metagenomic population (Kovaka et al., 2020; Payne et al., 2020). For diagnosis of mycoses, fungal DNA would ideally be selected for while human DNA is rejected; however, this is yet to be explored. The improvement of host to pathogen DNA ratio would greatly increase the amount of relevant data for clinical diagnosis and further work is required to implement these methods to routine diagnostics.

Lack of Reference Genomes

The major appeal of long-read sequencing in a shotgun approach is the vast genomic information that can be gained from the data generated. In metagenomic studies, genomic information from throughout the genome is generated; however, there is a lack of reference genomes for most of the pathogenic fungal species (**Figure 4**) and as such a proportion of data generated cannot be mapped to an appropriate reference genome either in FungiDB[5] (Basenko et al., 2018) and GenBank[6] (Sayers et al., 2020).

The basis for accurate genomic-based identification is the availability of high-quality, curated reference whole-genome databases. It has been demonstrated that inaccurate reference sequences have caused miss-mapping of the generated sequence reads leading to clear misidentification that potentially causes confusion surrounding diagnosis in clinical practice (Bialasiewicz et al., 2019; Irinyi et al., 2020). In addition, it is causing an unproductive strain on bioinformatic tools used and lengthening the time for identification. Diagnostic whole-genome database entries must have accurate sequences, metadata, and phylogenetic coverage to prevent false positive and negative identifications.

It is predicted that with the implementation of a reliable and accurate whole-genome reference database, the accuracy of fungal identification will dramatically improve (**Figure 5**). The principle is demonstrated through the developing taxonomy of the *Fusarium solani* species complex. A previous study examined four loci of these species and proposed the change in nomenclature of the *F. solani* species complex to be placed in the *Neocosmospora* genus (Sandoval-Denis et al., 2019). However, a recent study has analyzed 19 loci and has maintained that the members of the *F. solani* species complex should remain in the genus *Fusarium* (Geiser et al., 2020; O'Donnell et al., 2020). As such, the increased loci analyzed generated more data and provided more insight into the taxonomy of the *F. solani* species complex and if whole-genome data were available, the most accurate taxonomic information would be revealed, which if applied to all (pathogenic) fungal species

[5]https://fungidb.org/
[6]https://www.ncbi.nlm.nih.gov/genbank/

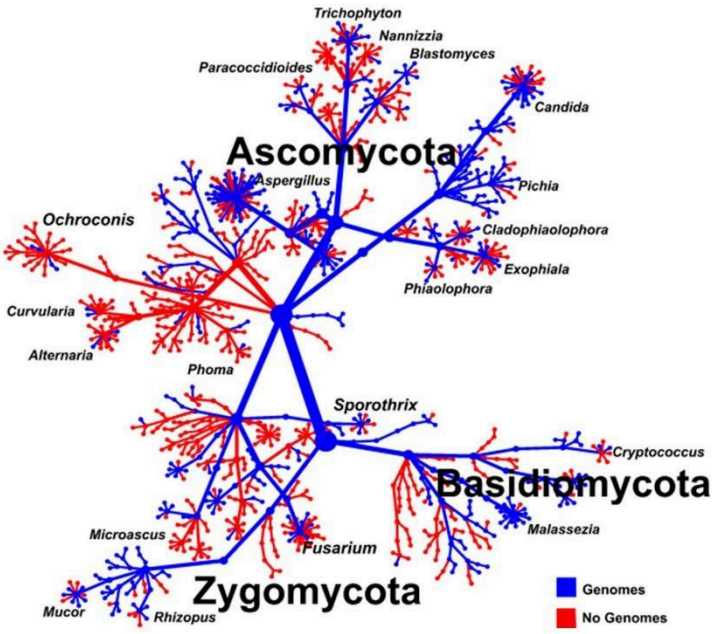


FIGURE 4 | Tree of the fungal Kingdom showing the major fungal classes and their human pathogenic species indicating species with (blue) and without (red) genomes.

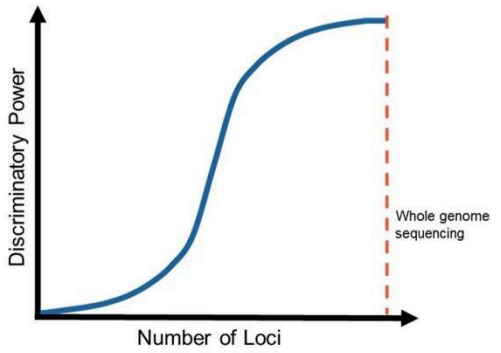


FIGURE 5 | Potential cumulative increase in discriminatory power with the addition of more loci or whole genomes being used for identification.

would greatly increase the discriminatory power of long-read sequencing in clinical diagnosis.

This highlights that global efforts are urgently needed to improve the relevant reference genome databases (FungiDB and GenBank) as well as the antifungal resistance database MARDy[7] (Nash et al., 2018), to be able to open up the full potential of the metagenomic data generated by long-read sequencers.

Bioinformatics

Currently, there is a wide variety of bioinformatics tools for the analysis of long-read sequences (Makalowski and Shabardina, 2020). Some such bioinformatics tools used for species identification in metagenomic studies are "What's In My Pot" (Juul et al., 2015), SUPRI real-time (Gu et al., 2021), MetaMaps (Dilthey et al., 2019), BugSeq (Fan et al., 2021), and NGSpeciesID (Vasiljevic et al., 2021).

For use in a clinical setting, a simple bioinformatics workflow would be necessary for practical implementation into routine diagnosis as it would be best to avoid requiring highly specialized personnel for analysis. A study comparing the ability of different bioinformatics tools to identify a wide array of pathogens from long-read sequencing dataset would be necessary to determine the most suitable bioinformatics tools.

Data interpretation is a major issue in metagenomics and metabarcoding analysis. Both methods generate large amounts of data and for the purpose of clinical diagnosis, the interpretation can be complex. When samples are taken from sterile sites, the expected sequencing results would consist of predominantly *Homo sapiens* reads and reads assigned the causative pathogen. However, if any contamination was to occur and reads were assigned to multiple microbes, this would complicate identification of the causative pathogen. This issue is magnified when patient samples are taken from non-sterile sites. Sequencing results will have reads assigned to multiple microbes and although it would be intuitive to assume the microbe with the highest abundance would be the causative pathogen, this is not always the case as has been observed in Irinyi et al. (2020). To overcome this, the application of machine-leaning algorithms helping to distinguish true pathogens from common colonizers or environmental contaminants may one of the ways forward. The large proportion of *H. sapiens* reads also complicates data interpretation as percentages of reads assigned to microorganisms is comparatively small. However, if bioinformatics tools were designed to remove host reads, microorganism abundance can be more easily compared and may further illuminate the causative pathogen.

[7]http://mardy.dide.ic.ac.uk/

THE FUTURE OF LONG-READ SEQUENCING AND FUNGAL DIAGNOSIS

The introduction of long-read sequencers from PacBio and ONT has allowed sequencing of genetic material in a way previously not possible. The first applications of long-read sequencing to identify agents of infectious disease has demonstrated its potential in clinical diagnosis.

The current state of DNA extraction technologies, the lack of comprehensive reference databases, and simple bioinformation pipelines requires further development before long-read sequencers can be implemented in routine diagnosis. Additionally, the identification of multiple microbes within a clinical specimen may not directly lead to a specific diagnosis, but in conjunction with the patient's presentation and further testing should seriously be considered as a way to achieve a fast and reliable diagnosis. Further advancements in the clinical diagnostic metagenomics workflow must be made to make its implementation in routine clinical diagnostic feasible. With the rapid decline of sequencing costs and the exponential improvement in sequencing capabilities, long-read sequencing is an important new technology which may eventually reduce the impact of pathogenic fungi on human health, by drastically reducing the turnaround time for diagnosis, supporting better treatment choices, reducing morbidity and mortality, and massively reduce associated healthcare costs, through its use in clinical diagnosis.

AUTHOR CONTRIBUTIONS

WM, LI, and BS conceived the review. WM coordinated and supervised the review. MH, LI, YH, BS, and WM wrote and corrected the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by a National Health and Medical Research Council of Australia (NH&MRC) grant (no. APP1121936) to WM.

REFERENCES

Angeletti, S. (2017). Matrix assisted laser desorption time of flight mass spectrometry (MALDI-TOF MS) in clinical microbiology. *J. Microbiol. Methods* 138, 20–29. doi: 10.1016/j.mimet.2016.09.003

Ardui, S., Ameur, A., Vermeesch, J. R., and Hestand, M. S. (2018). Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res.* 46, 2159–2168. doi: 10.1093/nar/gky066

Armanhi, J. S., de Souza, R. S., de Araujo, L. M., Okura, V. K., Mieczkowski, P., Imperial, J., et al. (2016). Multiplex amplicon sequencing for microbe identification in community-based culture collections. *Sci. Rep.* 6:29543. doi: 10.1038/srep29543

Ashikawa, S., Tarumoto, N., Imai, K., Sakai, J., Kodana, M., Kawamura, T., et al. (2018). Rapid identification of pathogens from positive blood culture bottles with the MinION nanopore sequencer. *J. Med. Microbiol.* 67, 1589–1595. doi: 10.1099/jmm.0.000855

Austin, B. (2017). The value of cultures to modern microbiology. *Antonie Van Leeuwenhoek* 110, 1247–1256. doi: 10.1007/s10482-017-0840-8

Basenko, E. Y., Pulman, J. A., Shanmugasundram, A., Harb, O. S., Crouch, K., Starns, D., et al. (2018). FungiDB: an integrated bioinformatic resource for fungi and oomycetes. *J. Fungi* 4, 1–28. doi: 10.3390/jof4010039

Bergfors, A., Leenheer, D. Bergqvist, A., Ameur, A., and Lennerstrand, J. (2016). Analysis of hepatitis C NS5A resistance associated polymorphisms using ultra deep single molecule real time (SMRT) sequencing. *Antivir. Res.* 126, 81–89. doi: 10.1016/j.antiviral.2015.12.005

Bharti, R., and Grimm, D. G. (2021). Current challenges and best-practice protocols for microbiome analysis. *Brief. Bioinform.* 22, 178–193. doi: 10.1093/bib/bbz155

Bialasiewicz, S., Duarte, T. P., Nguyen, S. H., Sukumaran, V., Stewart, A., Appleton, S., et al. (2019). Rapid diagnosis of *Capnocytophaga canimorsus* septic shock in an immunocompetent individual using real-time Nanopore sequencing: a case report. *BMC Infect. Dis.* 19:660. doi: 10.1186/s12879-019-4173-2

Brandt, M. E., Padhye, A. A., Mayer, L. W., and Holloway, B. P. (1998). Utility of random amplified polymorphic DNA PCR and TaqMan automated detection in molecular identification of *Aspergillus fumigatus*. *J. Clin. Microbiol.* 36, 2057–2062. doi: 10.1128/JCM.36.7.2057-2062.1998

Bull, R. A., Eltahla, A. A., Rodrigo, C., Koekkoek, S. M., Walker, M., Pirozyan, M. R., et al. (2016). A method for near full-length amplification and sequencing for six hepatitis C virus genotypes. *BMC Genomics* 17:247. doi: 10.1186/s12864-016-2575-8

Carvalho, A., Costa-De-Oliveira, S., Martins, M. L., Pina-Vaz, C., Rodrigues, A. G., Ludovico, P., et al. (2007). Multiplex PCR identification of eight clinically relevant *Candida* species. *Med. Mycol.* 45, 619–627. doi: 10.1080/13693780701501787

Castro-Wallace, S. L., Chiu, C. Y., John, K. K., Stahl, S. E., Rubins, K. H., McIntyre, A. B. R., et al. (2017). Nanopore DNA sequencing and genome assembly on the international space station. *Sci. Rep.* 7:18022. doi: 10.1038/s41598-017-18364-0

Chan, W. S., Au, C. H., Leung, S. M., Ho, D. N., Wong, E. Y. L., To, M. Y., et al. (2020). Potential utility of targeted nanopore sequencing for improving etiologic diagnosis of bacterial and fungal respiratory infection. *Diagn. Pathol.* 15:41. doi: 10.1186/s13000-020-00960-w

Charalampous, T., Kay, G. L., Richardson, H., Aydin, A., Baldan, R., Jeanes, C., et al. (2019). Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. *Nat. Biotechnol.* 37, 783–792. doi: 10.1038/s41587-019-0156-5

Chen, S. C.-A., Meyer, W., and Sorrell, T. C. (2014). *Cryptococcus gattii* infections. *Clin. Microbiol. Rev.* 27, 980–1024. doi: 10.1128/CMR.00126-13

Cheng, J., Hu, H., Kang, Y., Chen, W., Fang, W., Wang, K., et al. (2018). Identification of pathogens in culture-negative infective endocarditis cases by metagenomic analysis. *Ann. Clin. Microbiol. Antimicrob.* 17:43. doi: 10.1186/s12941-018-0294-5

Chin, C. S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* 10, 563–569. doi: 10.1038/nmeth.2474

Church, G., Deamer, D.W., Branton, D., Baldarelli, R., and Kasianowicz, J. (1998). *Characterization of Individual Polymer Molecules Based on Monomer-Interface Interactions.* US Patent No. 5795782. Available at: https://patents.google.com/patent/US5795782A/en (Accessed 23 February 2016).

Clancy, C. J., Pappas, P. G., Vazquez, J., Judson, M. A., Kontoyiannis, D. P., Thompson, G. R. 3rd, et al. (2018). Detecting infections rapidly and easily for candidemia trial, part 2 (DIRECT2): a prospective, multicenter study of the T2Candida panel. *Clin. Infect. Dis.* 66, 1678–1686. doi: 10.1093/cid/cix1095

Clark, A. E., Kaleta, E. J., Arora, A., and Wolk, D. M. (2013). Matrix-assisted laser desorption ionization-time of flight mass spectrometry: a fundamental shift in the routine practice of clinical microbiology. *Clin. Microbiol. Rev.* 26, 547–603. doi: 10.1128/CMR.00072-12

Cline, L. C., and Zak, D. R. (2015). Initial colonization, community assembly and ecosystem function: fungal colonist traits and litter biochemistry mediate decay rate. *Mol. Ecol.* 24, 5045–5058. doi: 10.1111/mec.13361

Colabella, C., Casagrande Pierantoni, D., Corte, L., Roscini, L., Conti, A., Bassetti, M., et al. (2021). Single strain high-depth NGS reveals high rDNA (ITS-LSU) variability in the four prevalent pathogenic species of the genus *Candida*. *Microorganisms* 9:302. doi: 10.3390/microorganisms9020302

Cornell, B. A., Braach-Maksvytis, V. L. B., King, L. G., Osman, P. D. J., Raguse, B., Wieczorek, L., et al. (1997). A biosensor that uses ion-channel switches. *Nature* 387, 580–583. doi: 10.1038/42432

Couto, N., Schuele, L., Raangs, E. C., Machado, M. P., Mendes, C. I., Jesus, T. F., et al. (2018). Critical steps in clinical shotgun metagenomics for the concomitant detection and typing of microbial pathogens. *Sci. Rep.* 8:13767. doi: 10.1038/s41598-018-31873-w

Cuomo, C. A., Shea, T., Yang, B., Rao, R., and Forche, A. (2017). Whole genome sequence of the heterozygous clinical isolate Candida krusei 81-B-5. *G3* 7, 2883–2889. doi: 10.1534/g3.117.043547

D'Andreano, S., Cusco, A., and Francino, O. (2021). Rapid and real-time identification of fungi up to species level with long amplicon nanopore sequencing from clinical samples. *Biol. Methods Protoc.* 6:bpaa026. doi: 10.1093/biomethods/bpaa026

de Heer, K., Gerritsen, M. G., Visser, C. E., and Leeflang, M. M. (2019). Galactomannan detection in broncho-alveolar lavage fluid for invasive aspergillosis in immunocompromised patients. *Cochrane Database Syst. Rev.* 5:CD012399. doi: 10.1002/14651858.CD012399.pub2

De Hoog, G., Guarro, J., Gené, J., Ahmed, S., Al-Hatmi, A., Figueras, M., et al. (2019). *Atlas of Clinical Fungi 2019*. Reus: Utrecht, The Netherlands.

De Maio, N., Shaw, L. P., Hubbard, A., George, S., Sanderson, N. D., Swann, J., et al. (2019). Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *Microb. Genom.* 5, 1–12. doi: 10.1099/mgen.0.000294

Deamer, D., Akeson, M., and Branton, D. (2016). Three decades of nanopore sequencing. *Nat. Biotechnol.* 34, 518–524. doi: 10.1038/nbt.3423

Dendis, M., Horvath, R., Michalek, J., Ruzicka, F., Grijalva, M., Bartos, M., et al. (2003). PCR-RFLP detection and species identification of fungal pathogens in patients with febrile neutropenia. *Clin. Microbiol. Infect.* 9, 1191–1202. doi: 10.1111/j.1469-0691.2003.00719.x

Deng, X., Achari, A., Federman, S., Yu, G., Somasekar, S., Bártolo, I., et al. (2020). Metagenomic sequencing with spiked primer enrichment for viral diagnostics and genomic surveillance. *Nat. Microbiol.* 5, 443–454. doi: 10.1038/s41564-019-0637-9

Dilernia, D. A., Chien, J.-T., Monaco, D. C., Brown, M. P., Ende, Z., Deymier, M. J., et al. (2015). Multiplexed highly-accurate DNA sequencing of closely-related HIV-1 variants using continuous long reads from single molecule, real-time sequencing. *Nucleic Acids Res.* 43:e129. doi: 10.1093/nar/gkv630

Dilthey, A. T., Jain, C., Koren, S., and Phillippy, A. M. (2019). Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps. *Nat. Commun.* 10:3066. doi: 10.1038/s41467-019-10934-2

Dudas, G., Carvalho, L. M., Bedford, T., Tatem, A. J., Baele, G., Faria, N. R., et al. (2017). Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature* 544, 309–315. doi: 10.1038/nature22040

Earl, J. P., Adappa, N. D., Krol, J., Bhat, A. S., Balashov, S., Ehrlich, R. L., et al. (2018). Species-level bacterial community profiling of the healthy sinonasal microbiome using pacific biosciences sequencing of full-length 16S rRNA genes. *Microbiome* 6:190. doi: 10.1186/s40168-018-0569-2

Edouard, S., Million, M., Bachar, D., Dubourg, G., Michelle, C., Ninove, L., et al. (2018). The nasopharyngeal microbiota in patients with viral respiratory tract infections is enriched in bacterial pathogens. *Eur. J. Clin. Microbiol. Infect. Dis.* 37, 1725–1733. doi: 10.1007/s10096-018-3305-8

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., et al. (2009). Real-time DNA sequencing from single polymerase molecules. *Science* 323, 133–138. doi: 10.1126/science.1162986

Eisenstein, M. (2012). Oxford nanopore announcement sets sequencing sector abuzz. *Nat. Biotechnol.* 30, 295–296. doi: 10.1038/nbt0412-295

El-Rami, F. E., Zielke, R. A., Wi, T., Sikora, A. E., and Unemo, M. (2019). Quantitative proteomics of the 2016 WHO *Neisseria gonorrhoeae* reference strains surveys vaccine candidates and antimicrobial resistance determinants. *Mol. Cell. Proteomics* 18, 127–150. doi: 10.1074/mcp.RA118.001125

Fan, J., Huang, S., and Chorlton, S. D. (2021). BugSeq: a highly accurate cloud platform for long-read metagenomic analyses. *BMC Bioinform.* 22:160. doi: 10.1186/s12859-021-04089-5

Faria, N. R., Kraemer, M. U. G., Hill, S. C., Goes de Jesus, J., Aguiar, R. S., Iani, F. C. M., et al. (2018). Genomic and epidemiological monitoring of yellow fever virus transmission potential. *Science* 361, 894–899. doi: 10.1126/science.aat7115

Faria, N. R., Quick, J., Claro, I. M., Theze, J., de Jesus, J. G., Giovanetti, M., et al. (2017). Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature* 546, 406–410. doi: 10.1038/nature22401

Feehan, J. M., Scheibel, K. E., Bourras, S., Underwood, W., Keller, B., and Somerville, S. C. (2017). Purification of high molecular weight genomic DNA from powdery mildew for long-read sequencing. *J. Vis. Exp.* e55463, 1–8. doi: 10.3791/55463

Feng, Z., Clemente, J. C., Wong, B., and Schadt, E. E. (2021). Detecting and phasing minor single-nucleotide variants from long-read sequencing data. *Nat. Commun.* 12:3032. doi: 10.1038/s41467-021-23289-4

Fenselau, C., and Demirev, P. A. (2001). Characterization of intact microorganisms by MALDI mass spectrometry. *Mass Spectrom. Rev.* 20, 157–171. doi: 10.1002/mas.10004

Forbes, J. D., Knox, N. C., Ronholm, J., Pagotto, F., and Reimer, A. (2017). Metagenomics: the next culture-independent game changer. *Front. Microbiol.* 8:1069. doi: 10.3389/fmicb.2017.01069

Ganley, A. R., and Kobayashi, T. (2007). Highly efficient concerted evolution in the ribosomal DNA repeats: total rDNA repeat variation revealed by whole-genome shotgun sequence data. *Genome Res.* 17, 184–191. doi: 10.1101/gr.5457707

Gargis, A. S., Cherney, B., Conley, A. B., McLaughlin, H. P., and Sue, D. (2019). Rapid detection of genetic engineering, structural variation, and antimicrobial resistance markers in bacterial biothreat pathogens by nanopore sequencing. *Sci. Rep.* 9:13501. doi: 10.1038/s41598-019-49700-1

Garrido-Cardenas, J. A., and Manzano-Agugliaro, F. (2017). The metagenomics worldwide research. *Curr. Genet.* 63, 819–829. doi: 10.1007/s00294-017-0693-8

Geiser, D. M., Al-Hatmi, A., Aoki, T., Arie, T., Balmas, V., Barnes, I., et al. (2020). Phylogenomic analysis of a 55.1 kb 19-gene dataset resolves a monophyletic *Fusarium* that includes the *Fusarium solani* species complex. *Phytopathology* 111, 1064–1079. doi: 10.1094/PHYTO-08-20-0330-LE

Goldstein, E. J., Tyrrell, K. L., and Citron, D. M. (2015). *Lactobacillus* species: taxonomic complexity and controversial susceptibilities. *Clin. Infect. Dis.* 60(*Suppl. 2*), S98–S107. doi: 10.1093/cid/civ072

Gonzalez-Escalona, N., Allard, M. A., Brown, E. W., Sharma, S., and Hoffmann, M. (2019). Nanopore sequencing for fast determination of plasmids, phages, virulence markers, and antimicrobial resistance genes in Shiga toxin-producing *Escherichia coli*. *PLoS One* 14:e0220494. doi: 10.1371/journal.pone.0220494

Gorrie, C. L., Mirceta, M., Wick, R. R., Judd, L. M., Wyres, K. L., Thomson, N. R., et al. (2018). Antimicrobial-resistant *Klebsiella pneumoniae* carriage and infection in specialized geriatric care wards linked to acquisition in the referring hospital. *Clin. Infect. Dis.* 67, 161–170. doi: 10.1093/cid/ciy027

Greig, D. R., Jenkins, C., Gharbia, S., and Dallman, T. J. (2019). Comparison of single-nucleotide variants identified by illumina and oxford nanopore technologies in the context of a potential outbreak of Shiga toxin-producing *Escherichia coli*. *GigaScience* 8, 1–12. doi: 10.1093/gigascience/giz104

Greninger, A. L., Naccache, S. N., Federman, S., Yu, G., Mbala, P., Bres, V., et al. (2015). Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Med.* 7:99. doi: 10.1186/s13073-015-0220-9

Gu, W., Deng, X., Lee, M., Sucu, Y. D., Arevalo, S., Stryke, D., et al. (2021). Rapid pathogen detection by metagenomic next-generation sequencing of infected body fluids. *Nat. Med.* 27, 115–124. doi: 10.1038/s41591-020-1105-z

Guarner, J., and Brandt, M. E. (2011). Histopathologic diagnosis of fungal infections in the 21st century. *Clin. Microbiol. Rev.* 24, 247–280. doi: 10.1128/CMR.00053-10

Gustafson, E., Bakotic, W., Bennett, L., Page, L., and McCarthy, L. (2019). DNA-based detection for onychomycosis correlates better to histopathology than does fungal culture. *Dermatol. Online J.* 25, 1–9.

Hahn, A., Sanyal, A., Perez, G. F., Colberg-Poley, A. M., Campos, J., Rose, M. C., et al. (2016). Different next generation sequencing platforms produce different microbial profiles and diversity in cystic fibrosis sputum. *J. Microbiol. Methods* 130, 95–99. doi: 10.1016/j.mimet.2016.09.002

Hamad, I., Abou Abdallah, R., Ravaux, I., Mokhtari, S., Tissot-Dupont, H., Michelle, C., et al. (2018). Metabarcoding analysis of eukaryotic microbiota in the gut of HIV-infected patients. *PLoS One* 13:e0191913. doi: 10.1371/journal.pone.0191913

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., and Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* 5, R245–R249. doi: 10.1016/S1074-5521(98)90108-9

Heather, J. M., and Chain, B. (2016). The sequence of sequencers: the history of sequencing DNA. *Genomics* 107, 1–8. doi: 10.1016/j.ygeno.2015.11.003

Hebert, P. D., Cywinska, A., Ball, S. L., and Dewaard, J. R. (2003). Biological identifications through DNA barcodes. *Proc. Biol. Sci.* 270, 313–321. doi: 10.1098/rspb.2002.2218

Heeger, F., Bourne, E. C., Baschien, C., Yurkov, A., Bunk, B., Sproer, C., et al. (2018). Long-read DNA metabarcoding of ribosomal RNA in the analysis of fungi from aquatic environments. *Mol. Ecol. Resour.* 18, 1500–1514. doi: 10.1111/1755-0998.12937

Helmersen, K., and Aamot, H. V. (2020). DNA extraction of microbial DNA directly from infected tissue: an optimized protocol for use in nanopore sequencing. *Sci. Rep.* 10:2985. doi: 10.1038/s41598-020-59957-6

Heravi, F. S., Zakrzewski, M., Vickery, K., and Hu, H. (2020). Host DNA depletion efficiency of microbiome DNA enrichment methods in infected tissue samples. *J. Microbiol. Methods* 170:105856. doi: 10.1016/j.mimet.2020.105856

Ho, C. K., Raghwani, J., Koekkoek, S., Liang, R. H., Van der Meer, J. T., Van Der Valk, M., et al. (2017). Characterization of hepatitis C virus (HCV) envelope diversification from acute to chronic infection within a sexually transmitted HCV cluster by using single-molecule, real-time sequencing. *J. Virol.* 91, 1–14. doi: 10.1128/JVI.02262-16

Hoenen, T., Groseth, A., Rosenke, K., Fischer, R. J., Hoenen, A., Judson, S. D., et al. (2016). Nanopore sequencing as a rapidly deployable Ebola outbreak tool. *Emerg. Infect. Dis.* 22:331. doi: 10.3201/eid2202.151796

Hu, K., Huang, N., Zou, Y., Liao, X., and Wang, J. (2021). MultiNanopolish: refined grouping method for reducing redundant calculations in nanopolish. *Bioinformatics* 37, 2757–2760. doi: 10.1093/bioinformatics/btab078

Huang, D. W., Raley, C., Jiang, M. K., Zheng, X., Liang, D., Rehman, M. T., et al. (2016). Towards better precision medicine: PacBio single-molecule long reads resolve the interpretation of HIV drug resistant mutation profiles at explicit quasispecies (haplotype) level. *J. Data Min. Genom. Proteom.* 7, 1–16. doi: 10.4172/2153-0602.1000182

Illumina (2021). Focused Power on the MiSeq System [Online]. Available: https://sapac.illumina.com/systems/sequencing-platforms/miseq.html (Accessed April 11, 2021).

Im, K., Mareninov, S., Diaz, M. F. P., and Yong, W. H. (2019). An introduction to performing immunofluorescence staining. *Methods Mol. Biol.* 1897, 299–311. doi: 10.1007/978-1-4939-8935-5_26

Imai, K., Nemoto, R., Kodana, M., Tarumoto, N., Sakai, J., Kawamura, T., et al. (2020). Rapid and accurate species identification of mitis group streptococci using the MinION nanopore sequencer. *Front. Cell. Infect. Microbiol.* 10:11. doi: 10.3389/fcimb.2020.00011

Inglis, P. W., Pappas, M. C. R., Resende, L. V., and Grattapaglia, D. (2018). Fast and inexpensive protocols for consistent extraction of high quality DNA and RNA from challenging plant and fungal samples for high-throughput SNP genotyping and sequencing applications. *PLoS One* 13:e0206085. doi: 10.1371/journal.pone.0206085

Irinyi, L., Hu, Y., Hoang, M. T. V., Pasic, L., Halliday, C., Jayawardena, M., et al. (2020). Long-read sequencing based clinical metagenomics for the detection and confirmation of *Pneumocystis jirovecii* directly from clinical specimens: a paradigm shift in mycological diagnostics. *Med. Mycol.* 58, 650–660. doi: 10.1093/mmy/myz109

Irinyi, L., Lackner, M., de Hoog, G. S., and Meyer, W. (2016). DNA barcoding of fungi causing infections in humans and animals. *Fungal Biol.* 120, 125–136. doi: 10.1016/j.funbio.2015.04.007

Irinyi, L., Serena, C., Garcia-Hermoso, D., Arabatzis, M., Desnos-Ollivier, M., Vu, D., et al. (2015). International Society of Human and Animal Mycology (ISHAM)-ITS reference DNA barcoding database—the quality controlled standard tool for routine identification of human and animal pathogenic fungi. *Med. Mycol.* 53, 313–337. doi: 10.1093/mmy/myv008

Jain, M., Olsen, H. E., Paten, B., and Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* 17:239. doi: 10.1186/s13059-016-1103-0

James, T. Y., Marino, J. A., Perfecto, I., and Vandermeer, J. (2016). Identification of putative coffee rust mycoparasites via single-molecule DNA sequencing of infected pustules. *J. Appl. Environ. Microbiol.* 82, 631–639. doi: 10.1128/AEM.02639-15

Janda, J. M., and Abbott, S. L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *J. Clin. Microbiol.* 45, 2761–2764. doi: 10.1128/JCM.01228-07

Jumpponen, A. (2003). Soil fungal community assembly in a primary successional glacier forefront ecosystem as inferred from rDNA sequence analyses. *New Phytol.* 158, 569–578. doi: 10.1046/j.1469-8137.2003.00767.x

Juul, S., Izquierdo, F., Hurst, A., Dai, X., Wright, A., Kulesha, E., et al. (2015). What's in my pot? Real-time species identification on the MinION™. bioRxiv. doi: 10.1101/030742

Kafetzopoulou, L. E., Efthymiadis, K., Lewandowski, K., Crook, A., Carter, D., Osborne, J., et al. (2018). Assessment of metagenomic Nanopore and Illumina sequencing for recovering whole genome sequences of chikungunya and dengue viruses directly from clinical samples. *Euro Surveill.* 23, 1–13. doi: 10.2807/1560-7917.ES.2018.23.50.1800228

Kafetzopoulou, L., Pullan, S., Lemey, P., Suchard, M., Ehichioya, D., Pahlmann, M., et al. (2019). Metagenomic sequencing at the epicenter of the Nigeria 2018 Lassa fever outbreak. *Science* 363, 74–77. doi: 10.1126/science.aau9343

Kai, S., Matsuo, Y., Nakagawa, S., Kryukov, K., Matsukawa, S., Tanaka, H., et al. (2019). Rapid bacterial identification by direct PCR amplification of 16S rRNA genes using the MinION™ nanopore sequencer. *FEBS Open Bio* 9, 548–557. doi: 10.1002/2211-5463.12590

Kanamori, H., Parobek, C. M., Weber, D. J., van Duin, D., Rutala, W. A., Cairns, B. A., et al. (2016). Next-generation sequencing and comparative analysis of sequential outbreaks caused by multidrug-resistant *Acinetobacter baumannii* at a large academic burn center. *Antimicrob. Agents Chemother.* 60, 1249–1257. doi: 10.1128/AAC.02014-15

Karageorgopoulos, D. E., Vouloumanou, E. K., Ntziora, F., Michalopoulos, A., Rafailidis, P. I., and Falagas, M. E. (2011). Beta-D-glucan assay for the diagnosis of invasive fungal infections: a meta-analysis. *Clin. Infect. Dis.* 52, 750–770. doi: 10.1093/cid/ciq206

Kasianowicz, J. J., Brandin, E., Branton, D., and Deamer, D. W. (1996). Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci. U. S. A.* 93, 13770–13773. doi: 10.1073/pnas.93.24.13770

Kidd, S. E., Chen, S. C., Meyer, W., and Halliday, C. L. (2019). A new age in molecular diagnostics for invasive fungal disease: are we ready? *Front. Microbiol.* 10:2903. doi: 10.3389/fmicb.2019.02903

Klingspor, L., and Jalal, S. (2006). Molecular detection and identification of *Candida* and *Aspergillus* spp. from clinical samples using real-time PCR. *Clin. Microbiol. Infect.* 12, 745–753. doi: 10.1111/j.1469-0691.2006.01498.x

Ko, S. H., Bayat Mokhtari, E., Mudvari, P., Stein, S., Stringham, C. D., Wagner, D., et al. (2021). High-throughput, single-copy sequencing reveals SARS-CoV-2 spike variants coincident with mounting humoral immunity during acute COVID-19. *PLoS Pathog.* 17:e1009431. doi: 10.1371/journal.ppat.1009431

Koljalg, U., Larsson, K. H., Abarenkov, K., Nilsson, R. H., Alexander, I. J., Eberhardt, U., et al. (2005). UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. *New Phytol.* 166, 1063–1068. doi: 10.1111/j.1469-8137.2005.01376.x

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 27, 722–736. doi: 10.1101/gr.215087.116

Kovaka, S., Fan, Y., Ni, B., Timp, W., and Schatz, M. C. (2020). Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. *Nat. Biotechnol.* 29, 431–441. doi: 10.1038/s41587-020-0731-9

Kulik, T., Fordoński, G., Pszczółkowska, A., Płodzień, K., and Łapiński, M. (2004). Development of PCR assay based on ITS2 rDNA polymorphism for the detection and differentiation of *Fusarium sporotrichioides*. *FEMS Microbiol. Lett.* 239, 181–186. doi: 10.1016/j.femsle.2004.08.037

Kurtzman, C., Fell, J. W., and Boekhout, T. (2011). *The Yeasts: A Taxonomic Study.* Amsterdam, Netherlands: Elsevier.

Kyaschenko, J., Clemmensen, K. E., Hagenbo, A., Karltun, E., and Lindahl, B. D. (2017). Shift in fungal communities and associated enzyme activities along an age gradient of managed Pinus sylvestris stands. *ISME J.* 11, 863–874. doi: 10.1038/ismej.2016.184

Lackner, M., de Hoog, G. S., Yang, L., Moreno, L. F., Ahmed, S. A., Andreas, F., et al. (2014). Proposed nomenclature for *Pseudallescheria*, *Scedosporium* and related genera. *Fungal Divers.* 67, 1–10. doi: 10.1007/s13225-014-0295-4

Laudadio, I., Fulci, V., Palone, F., Stronati, L., Cucchiara, S., and Carissimi, C. (2018). Quantitative assessment of shotgun metagenomics and 16S rDNA amplicon sequencing in the study of human gut microbiome. *OMICS* 22, 248–254. doi: 10.1089/omi.2018.0013

Leggett, R. M., and Clark, M. D. (2017). A world of opportunities with nanopore sequencing. *J. Exp. Bot.* 68, 5419–5429. doi: 10.1093/jxb/erx289

Lewandowski, K., Xu, Y., Pullan, S. T., Lumley, S. F., Foster, D., Sanderson, N., et al. (2019). Metagenomic nanopore sequencing of influenza virus direct from clinical respiratory samples. *J. Clin. Microbiol.* 58, 1–15. doi: 10.1128/JCM.00963-19

Li, Y., He, X. Z., Li, M. H., Li, B., Yang, M. J., Xie, Y., et al. (2020). Comparison of third-generation sequencing approaches to identify viral pathogens under public health emergency conditions. *Virus Genes* 56, 288–297. doi: 10.1007/s11262-020-01746-4

Lieckfeldt, E., Meyer, W., and Borner, T. (1993). Rapid identification and differentiation of yeasts by DNA and PCR fingerprinting. *J. Basic Microbiol.* 33, 413–425. doi: 10.1002/jobm.3620330609

Lindsley, M. D., Hurst, S. F., Iqbal, N. J., and Morrison, C. J. (2001). Rapid identification of dimorphic and yeast-like fungal pathogens using specific DNA probes. *J. Clin. Microbiol.* 39, 3505–3511. doi: 10.1128/JCM.39.10.3505-3511.2001

Logsdon, G. A., Vollger, M. R., and Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* 21, 597–614. doi: 10.1038/s41576-020-0236-x

Loit, K., Adamson, K., Bahram, M., Puusepp, R., Anslan, S., Kiiker, R., et al. (2019). Relative performance of MinION (Oxford Nanopore technologies) versus sequel (Pacific biosciences) third-generation sequencing instruments in identification of agricultural and Forest fungal pathogens. *J. Appl. Environ. Microbiol.* 85, 1–20. doi: 10.1128/AEM.01368-19

Loman, N. J., Quick, J., and Simpson, J. T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* 12, 733–735. doi: 10.1038/nmeth.3444

Lu, H., Giordano, F., and Ning, Z. (2016). Oxford nanopore MinION sequencing and genome assembly. *Genom. Proteom. Bioinform.* 14, 265–279. doi: 10.1016/j.gpb.2016.05.004

Lui, W. Y., Yuen, C. K., Li, C., Wong, W. M., Lui, P. Y., Lin, C. H., et al. (2019). SMRT sequencing revealed the diversity and characteristics of defective interfering RNAs in influenza A (H7N9) virus infection. *Emerg. Microbes Infect.* 8, 662–674. doi: 10.1080/22221751.2019.1611346

Luo, R., Zimin, A., Workman, R., Fan, Y., Pertea, G., Grossman, N., et al. (2017). First draft genome sequence of the pathogenic fungus lomentospora prolificans (formerly scedosporium prolificans). *G3* 7, 3831–3836. doi: 10.1534/g3.117.300107

Mafune, K. K., Godfrey, B. J., Vogt, D. J., and Vogt, K. A. (2020). A rapid approach to profiling diverse fungal communities using the MinION nanopore sequencer. *BioTechniques* 68, 72–78. doi: 10.2144/btn-2019-0072

Makalowski, W., and Shabardina, V. (2020). Bioinformatics of nanopore sequencing. *J. Hum. Genet.* 65, 61–67. doi: 10.1038/s10038-019-0659-4

Mannarelli, B. M., and Kurtzman, C. P. (1998). Rapid identification of *Candida albicans* and other human pathogenic yeasts by using short oligonucleotides

in a PCR. *J. Clin. Microbiol.* 36, 1634–1641. doi: 10.1128/JCM.36.6.1634-1641.1998

Mantere, T., Kersten, S., and Hoischen, A. (2019). Long-read sequencing emerging in medical genetics. *Front. Genet.* 10:426. doi: 10.3389/fgene.2019.00426

Marotz, C. A., Sanders, J. G., Zuniga, C., Zaramela, L. S., Knight, R., and Zengler, K. (2018). Improving saliva shotgun metagenomics by chemical host DNA depletion. *Microbiome* 6:42. doi: 10.1186/s40168-018-0426-3

Martin, C., Roberts, D., van der Weide, M., Rossau, R., Jannes, G., Smith, T., et al. (2000). Development of a PCR-based line probe assay for identification of fungal pathogens. *J. Clin. Microbiol.* 38, 3735–3742. doi: 10.1128/JCM.38.10.3735-3742.2000

Meredith, L. W., Hamilton, W. L., Warne, B., Houldcroft, C. J., Hosmillo, M., Jahun, A. S., et al. (2020). Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. *Lancet Infect. Dis.* 20, 1263–1272. doi: 10.1016/S1473-3099(20)30562-4

Meyer, W., Irinyi, L., Hoang, M. T. V., Robert, V., Garcia-Hermoso, D., Desnos-Ollivier, M., et al. (2019). Database establishment for the secondary fungal DNA barcode *translational elongation factor 1α* (*TEF1α*). *Genome* 62, 160–169. doi: 10.1139/gen-2018-0083

Miller, S., and Chiu, C. (2018). "Metagenomic next-generation sequencing for pathogen detection and identification," in *Advanced Techniques in Diagnostic Microbiology: Volume 2: Applications.* eds. Y.-W. Tang and C. W. Stratton (Cham: Springer International Publishing), 617–632.

Mobasherizadeh, S., Shojaei, H., Havaei, S. A., Mostafavizadeh, K., Davoodabadi, F., Khorvash, F., et al. (2016). Application of the random amplified polymorphic DNA (RAPD) fingerprinting to analyze genetic variation in community associated-methicillin resistant *staphylococcus Aureus* (CA-MRSA) isolates in Iran. *Global J. Health Sci.* 8:53822. doi: 10.5539/gjhs.v8n8p185

Moon, J., Kim, N., Kim, T. J., Jun, J. S., Lee, H. S., Shin, H. R., et al. (2019). Rapid diagnosis of bacterial meningitis by nanopore 16S amplicon sequencing: a pilot study. *Int. J. Med. Microbiol.* 309:151338. doi: 10.1016/j.ijmm.2019.151338

Morand, S. C., Bertignac, M., Iltis, A., Kolder, I., Pirovano, W., Jourdain, R., et al. (2019). Complete genome sequence of *Malassezia restricta* CBS 7877, an opportunist pathogen involved in dandruff and seborrheic dermatitis. *Microbiol. Resour. Announc.* 8, 1–3. doi: 10.1128/MRA.01543-18

Morrison, G. A., Fu, J., Lee, G. C., Wiederhold, N. P., Canete-Gibas, C. F., Bunnik, E. M., et al. (2020). Nanopore sequencing of the fungal intergenic spacer sequence as a potential rapid diagnostic assay. *J. Clin. Microbiol.* 58, 1–16. doi: 10.1128/JCM.01972-20

Motooka, D., Fujimoto, K., Tanaka, R., Yaguchi, T., Gotoh, K., Maeda, Y., et al. (2017). Fungal ITS1 deep-sequencing strategies to reconstruct the composition of a 26-species community and evaluation of the gut mycobiota of healthy japanese individuals. *Front. Microbiol.* 8:238. doi: 10.3389/fmicb.2017.00238

Mylonakis, E., Clancy, C. J., Ostrosky-Zeichner, L., Garey, K. W., Alangaden, G. J., Vazquez, J. A., et al. (2015). T2 magnetic resonance assay for the rapid diagnosis of candidemia in whole blood: a clinical trial. *Clin. Infect. Dis.* 60, 892–899. doi: 10.1093/cid/ciu959

Nakagawa, S., Inoue, S., Kryukov, K., Yamagishi, J., Ohno, A., Hayashida, K., et al. (2019). Rapid sequencing-based diagnosis of infectious bacterial species from meningitis patients in Zambia. *Clin. Transl. Immunol.* 8:e01087. doi: 10.1002/cti2.1087

Nakano, K., Shiroma, A., Shimoji, M., Tamotsu, H., Ashimine, N., Ohki, S., et al. (2017). Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. *Hum. Cell* 30, 149–161. doi: 10.1007/s13577-017-0168-8

Nash, A., Sewell, T., Farrer, R. A., Abdolrasouli, A., Shelton, J. M. G., Fisher, M. C., et al. (2018). MARDy: mycology antifungal resistance database. *Bioinformatics* 34, 3233–3234. doi: 10.1093/bioinformatics/bty321

Neely, L. A., Audeh, M., Phung, N. A., Min, M., Suchocki, A., Plourde, D., et al. (2013). T2 magnetic resonance enables nanoparticle-mediated rapid detection of candidemia in whole blood. *Sci. Transl. Med.* 5:182ra154. doi: 10.1126/scitranslmed.3005377

Nilsson, R. H., Ryberg, M., Abarenkov, K., Sjökvist, E., and Kristiansson, E. (2009). The ITS region as a target for characterization of fungal communities using emerging sequencing technologies. *FEMS Microbiol. Lett.* 296, 97–101. doi: 10.1111/j.1574-6968.2009.01618.x

Ocwieja, K. E., Sherrill-Mix, S., Mukherjee, R., Custers-Allen, R., David, P., Brown, M., et al. (2012). Dynamic regulation of HIV-1 mRNA populations analyzed by single-molecule enrichment and long-read sequencing. *Nucleic Acids Res.* 40, 10345–10355. doi: 10.1093/nar/gks753

O'Donnell, K., Al-Hatmi, A. M. S., Aoki, T., Brankovics, B., Cano-Lira, J. F., Coleman, J. J., et al. (2020). No to neocosmospora: phylogenomic and practical reasons for continued inclusion of the *Fusarium solani* species complex in the genus *Fusarium*. *mSphere* 5, 1–7. doi: 10.1128/mSphere.00810-20

O'Donnell, K., and Cigelnik, E. (1997). Two divergent intragenomic rDNA ITS2 types within a monophyletic lineage of the fungus *Fusarium* are nonorthologous. *Mol. Phylogenet. Evol.* 7, 103–116. doi: 10.1006/mpev.1996.0376

Oxford Nanopore Technologies (2019). Oxford Nanopore launches Flongle, for Rapid, Smaller DNA/RNA Sequencing Tests in Any Environment [Online]. Available: https://nanoporetech.com/about-us/news/oxford-nanopore-launches-flongle-rapid-smaller-dnarna-sequencing-tests-any (Accessed October 28, 2021).

Oxford Nanopore Technologies (2021). Products [Online]. Available: https://nanoporetech.com/products (Accessed October 28, 2021).

Pacific Biosciences (2020). Pacific Biosciences Launches the Sequel IIe System to Accelerate Adoption of Highly Accurate HiFi Sequencing [Online]. Available: https://www.pacb.com/press_releases/pacific-biosciences-launches-the-sequel-iie-system-to-accelerate-adoption-of-highly-accurate-hifi-sequencing/ (Accessed October 28, 2021).

Pacific Biosciences (2021a). PacBio Sequel Systems [Online]. Available: https://www.pacb.com/products-and-services/sequel-system/ (Accessed October 28, 2021).

Pacific Biosciences (2021b). What Can You Do with One SMRT Cell? [Online]. Available: https://www.pacb.com/wp-content/uploads/Application-Brochure-What-Can-You-Do-with-One-SMRT-Cell.pdf (Accessed October 28, 2021).

Panthee, S., Hamamoto, H., Ishijima, S. A., Paudel, A., and Sekimizu, K. (2018). Utilization of hybrid assembly approach to determine the genome of an opportunistic pathogenic fungus, *Candida albicans* TIMM 1768. *Genome Biol. Evol.* 10, 2017–2022. doi: 10.1093/gbe/evy166

Parker, J., Helmstetter, A. J., Devey, D., Wilkinson, T., and Papadopulos, A. S. (2017). Field-based species identification of closely-related plants using real-time nanopore sequencing. *Sci. Rep.* 7, 1–8. doi: 10.1038/s41598-017-08461-5

Patel, T. S., Kaakeh, R., Nagel, J. L., Newton, D. W., and Stevenson, J. G. (2017). Cost analysis of implementing matrix-assisted laser desorption ionization-time of flight mass spectrometry plus real-time antimicrobial stewardship intervention for bloodstream infections. *J. Clin. Microbiol.* 55, 60–67. doi: 10.1128/JCM.01452-16

Payne, A., Holmes, N., Clarke, T., Munro, R., Debebe, B. J., and Loose, M. (2020). Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nat. Biotechnol.* 35, 2193–2198. doi: 10.1038/s41587-020-00746-x

Payne, A., Holmes, N., Rakyan, V., and Loose, M. (2019). BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics* 35, 2193–2198. doi: 10.1093/bioinformatics/bty841

Pchelin, I. M., Azarov, D. V., Churina, M. A., Ryabinin, I. A., Vibornova, I. V., Apalko, S. V., et al. (2020). Whole genome sequence of first *Candida auris* strain, isolated in Russia. *Med. Mycol.* 58, 414–416. doi: 10.1093/mmy/myz078

Perfect, J. R. (2013). Fungal diagnosis: how do we do it and can we do better? *Curr. Med. Res. Opin.* 29(Suppl 4), 3–11. doi: 10.1185/03007995.2012.761134

Pfeiffer, C. D., Fine, J. P., and Safdar, N. (2006). Diagnosis of invasive aspergillosis using a galactomannan assay: a meta-analysis. *Clin. Infect. Dis.* 42, 1417–1727. doi: 10.1086/503427

Pollard, M. O., Gurdasani, D., Mentzer, A. J., Porter, T., and Sandhu, M. S. (2018). Long reads: their purpose and place. *Hum. Mol. Genet.* 27, R234–R241. doi: 10.1093/hmg/ddy177

Polz, M. F., and Cavanaugh, C. M. (1998). Bias in template-to-product ratios in multitemplate PCR. *J. Appl. Environ. Microbiol.* 64, 3724–3730. doi: 10.1128/AEM.64.10.3724-3730.1998

Pryce, T. M., Fau, P. S., Kay, I. D., and Coombs, G. W. (2003). Rapid identification of fungi by sequencing the ITS1 and ITS2 regions using an automated capillary electrophoresis system. *Med. Mycol.* 41, 369–381. doi: 10.1080/13693780310001600435

Purahong, W., Mapook, A., Wu, Y. T., and Chen, C. T. (2019). Characterization of the *Castanopsis carlesii* deadwood mycobiome by PacBio sequencing of the full-length fungal nuclear ribosomal internal transcribed spacer (ITS). *Front. Microbiol.* 10:983. doi: 10.3389/fmicb.2019.00983

Quick, J., Grubaugh, N. D., Pullan, S. T., Claro, I. M., Smith, A. D., Gangavarapu, K., et al. (2017). Multiplex PCR method for MinION and illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat. Protoc.* 12, 1261–1276. doi: 10.1038/nprot.2017.066

Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., et al. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature* 530, 228–232. doi: 10.1038/nature16996

Ratnasingham, S., and Hebert, P. D. (2007). Bold: the barcode of life data system. *Mol. Ecol. Notes* 7, 355–364. doi: 10.1111/j.1471-8286.2007.01678.x

Ravi, R. K., Walton, K., and Khosroheidari, M. (2018). MiSeq: a next generation sequencing platform for genomic analysis. *Methods Mol. Biol.* 1706, 223–232. doi: 10.1007/978-1-4939-7471-9_12

Rhoads, A., and Au, K. F. (2015). PacBio sequencing and its applications. *Genom. Proteom. Bioinform.* 13, 278–289. doi: 10.1016/j.gpb.2015.08.002

Rhodes, J., Abdolrasouli, A., Farrer, R. A., Cuomo, C. A., Aanensen, D. M., Armstrong-James, D., et al. (2018). Genomic epidemiology of the UK outbreak of the emerging human fungal pathogen *Candida auris*. *Emerg. Infect. Dis.* 7:43. doi: 10.1038/s41426-018-0045-x

Riesenfeld, C. S., Schloss, P. D., and Handelsman, J. (2004). Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet.* 38, 525–552. doi: 10.1146/annurev.genet.38.072902.091216

Romanelli, A. M., Sutton, D. A., Thompson, E. H., Rinaldi, M. G., and Wickes, B. L. (2010). Sequence-based identification of filamentous basidiomycetous fungi from clinical specimens: a cautionary note. *J. Clin. Microbiol.* 48, 741–752. doi: 10.1128/JCM.01948-09

Sakai, J., Tarumoto, N., Kodana, M., Ashikawa, S., Imai, K., Kawamura, T., et al. (2019). An identification protocol for ESBL-producing gram-negative bacteria bloodstream infections using a MinION nanopore sequencer. *J. Med. Microbiol.* 68, 1219–1226. doi: 10.1099/jmm.0.001024

Salmela, L., Walve, R., Rivals, E., and Ukkonen, E. (2017). Accurate self-correction of errors in long reads using de Bruijn graphs. *Bioinformatics* 33, 799–806. doi: 10.1093/bioinformatics/btw321

Salzberg, S. L., and Yorke, J. A. (2005). Beware of mis-assembled genomes. *Bioinformatics* 21, 4320–4321. doi: 10.1093/bioinformatics/bti769

Sandhu, G. S., Kline, B. C., Stockman, L., and Roberts, G. D. (1995). Molecular probes for diagnosis of fungal infections. *J. Clin. Microbiol.* 33, 2913–2919. doi: 10.1128/jcm.33.11.2913-2919.1995

Sandoval-Denis, M., Lombard, L., and Crous, P. W. (2019). Back to the roots: a reappraisal of *Neocosmospora*. *Persoonia* 43, 90–185. doi: 10.3767/persoonia.2019.43.04

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* 74, 5463–5467. doi: 10.1073/pnas.74.12.5463

Sayers, E. W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K. D., and Karsch-Mizrachi, I. (2020). GenBank. *Nucleic Acids Res.* 48, D84–D86. doi: 10.1093/nar/gkz956

Schelenz, S., Barnes, R. A., Barton, R. C., Cleverley, J. R., Lucas, S. B., Kibbler, C. C., et al. (2015). British society for medical mycology best practice recommendations for the diagnosis of serious fungal diseases. *Lancet Infect. Dis.* 15, 461–474. doi: 10.1016/S1473-3099(15)70006-X

Schmidt, K., Mwaigwisya, S., Crossman, L. C., Doumith, M., Munroe, D., Pires, C., et al. (2017). Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing. *J. Antimicrob. Chemother.* 72, 104–114. doi: 10.1093/jac/dkw397

Schoch, C. L., Robbertse, B., Robert, V., Vu, D., Cardinali, G., Irinyi, L., et al. (2014). Finding needles in haystacks: linking scientific names, reference specimens and molecular data for fungi. *Database* 2014, 1–21. doi: 10.1093/database/bau061

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., et al. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for fungi. *Proc. Natl. Acad. Sci. U. S. A.* 109, 6241–6246. doi: 10.1073/pnas.1117018109

Schultzhaus, Z., Cuomo, C. A., and Wang, Z. (2019). Genome sequence of the black yeast *Exophiala lecanii-corni*. *Microbiol. Resour. Announc.* 8, 1–2. doi: 10.1128/MRA.01709-18

Schwessinger, B., and Rathjen, J. P. (2017). Extraction of high molecular weight DNA from fungal rust spores for long read sequencing. *Methods Mol. Biol.* 1659, 49–57. doi: 10.1007/978-1-4939-7249-4_5

Shokralla, S., Gibson, J. F., Nikbakht, H., Janzen, D. H., Hallwachs, W., and Hajibabaei, M. (2014). Next-generation DNA barcoding: using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens. *Mol. Ecol. Resour.* 14, 892–901. doi: 10.1111/1755-0998.12236

Simon, U. K., and Weiss, M. (2008). Intragenomic variation of fungal ribosomal genes is higher than previously thought. *Mol. Biol. Evol.* 25, 2251–2254. doi: 10.1093/molbev/msn188

Springer, J., and Loffler, J. (2017). Genus- and species-specific PCR detection methods. *Methods Mol. Biol.* 1508, 267–279. doi: 10.1007/978-1-4939-6515-1_15

Stielow, J. B., Levesque, C. A., Seifert, K. A., Meyer, W., Irinyi, L., Smits, D., et al. (2015). One fungus, which genes? Development and assessment of universal primers for potential secondary fungal DNA barcodes. *Persoonia* 35:242. doi: 10.3767/003158515X689135

Takeda, H., Yamashita, T., Ueda, Y., and Sekine, A. (2019). Exploring the hepatitis C virus genome using single molecule real-time sequencing. *World J. Gastroenterol.* 25:4661. doi: 10.3748/wjg.v25.i32.4661

Taxt, A. M., Avershina, E., Frye, S. A., Naseer, U., and Ahmad, R. (2020). Rapid identification of pathogens, antibiotic resistance genes and plasmids in blood cultures by nanopore sequencing. *Sci. Rep.* 10:7622. doi: 10.1038/s41598-020-64616-x

Tedersoo, L., Albertsen, M., Anslan, S., and Callahan, B. (2021). Perspectives and benefits of high-throughput long-read sequencing in microbial ecology. *Appl. Environ. Microbiol.* 87:e0062621. doi: 10.1128/AEM.00626-21

Tedersoo, L., and Anslan, S. (2019). Towards PacBio-based pan-eukaryote metabarcoding using full-length ITS sequences. *Environ. Microbiol. Rep.* 11, 659–668. doi: 10.1111/1758-2229.12776

Tedersoo, L., Tooming-Klunderud, A., and Anslan, S. (2018). PacBio metabarcoding of fungi and other eukaryotes: errors, biases and perspectives. *New Phytol.* 217, 1370–1385. doi: 10.1111/nph.14776

Teng, J. L., Yeung, M. L., Chan, E., Jia, L., Lin, C. H., Huang, Y., et al. (2017). PacBio but not Illumina technology can achieve fast, accurate and complete closure of the high GC, complex *Burkholderia pseudomallei* two-chromosome genome. *Front. Microbiol.* 8:1448. doi: 10.3389/fmicb.2017.01448

Tia, T., Putaporntip, C., Kosuwin, R., Kongpolprom, N., Kawkitinarong, K., and Jongwutiwes, S. (2012). A highly sensitive novel PCR assay for detection of *Pneumocystis jirovecii* DNA in bronchoalveloar lavage specimens from immunocompromised patients. *Clin. Microbiol. Infect.* 18, 598–603. doi: 10.1111/j.1469-0691.2011.03656.x

Treangen, T. J., and Salzberg, S. L. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* 13, 36–46. doi: 10.1038/nrg3117

Tsang, C. C., Teng, J. L. L., Lau, S. K. P., and Woo, P. C. Y. (2021). Rapid genomic diagnosis of fungal infections in the age of next-generation sequencing. *J. Fungi* 7, 1–15. doi: 10.3390/jof7080636

Vale-Silva, L., Beaudoing, E., Tran, V. D. T., and Sanglard, D. (2017). Comparative genomics of two sequential *Candida glabrata* clinical isolates. *G3* 7, 2413–2426. doi: 10.1534/g3.117.042887

van Dijk, E. L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends Genet.* 30, 418–426. doi: 10.1016/j.tig.2014.07.001

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., and Thermes, C. (2018). The third revolution in sequencing technology. *Trends Genet.* 34, 666–681. doi: 10.1016/j.tig.2018.05.008

Vasiljevic, N., Lim, M., Humble, E., Seah, A., Kratzer, A., Morf, N. V., et al. (2021). Developmental validation of Oxford nanopore technology MinION sequence data and the NGSpeciesID bioinformatic pipeline for forensic genetic species identification. *Forensic Sci. Int. Genet.* 53:102493. doi: 10.1016/j.fsigen.2021.102493

Votintseva, A. A., Bradley, P., Pankhurst, L., Del Ojo Elias, C., Loose, M., Nilgiriwala, K., et al. (2017). Same-day diagnostic and surveillance data for tuberculosis via whole-genome sequencing of direct respiratory samples. *J. Clin. Microbiol.* 55, 1285–1298. doi: 10.1128/JCM.02483-16

Wagner, A., Chavez, V., Huang, R. S. P., Wahed, A., Actor, J. K., and Dasgupta, A. (2017). "Chapter 5: biochemical tests and staining techniques for microbial identification," in *Microbiology and Molecular Diagnosis in Pathology A Comprehensive Review for Board Preparation, Certification and Clinical Practice*. ed. T. Broderick and K. Padilla (Amsterdam, Netherlands: Elsevier), 61–73.

Wagner, J., Coupland, P., Browne, H. P., Lawley, T. D., Francis, S. C., and Parkhill, J. (2016). Evaluation of PacBio sequencing for full-length bacterial

16S rRNA gene classification. *BMC Microbiol.* 16:274. doi: 10.1186/s12866-016-0891-4

Walder, F., Schlaeppi, K., Wittwer, R., Held, A. Y., Vogelgsang, S., and van der Heijden, M. G. (2017). Community profiling of *Fusarium* in combination with other plant-associated fungi in different crop species using SMRT sequencing. *Front. Plant Sci.* 8:2019. doi: 10.3389/fpls.2017.02019

Wang, M., Fu, A., Hu, B., Shen, G., Liu, R., Zhao, W., et al. (2020). Same-day simultaneous diagnosis of bacterial and fungal infections in clinical practice by nanopore targeted sequencing. medRxiv. doi: 10.1101/2020.04.08.20057604

Wang, W. L., Xu, S. Y., Ren, Z. G., Tao, L., Jiang, J. W., and Zheng, S. S. (2015). Application of metagenomics in the human gut microbiome. *World J. Gastroenterol.* 21, 803–814. doi: 10.3748/wjg.v21.i3.803

Washington, N.L., Gangavarapu, K., Zeller, M., Bolze, A., Cirulli, E.T., Schiabor Barrett, K.M., et al. (2021). Genomic epidemiology identifies emergence and rapid transmission of SARS-CoV-2 B.1.1.7 in the United States. medRxiv. doi: 10.1101/2021.02.06.21251159

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P. C., Hall, R. J., Concepcion, G. T., et al. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* 37, 1155–1162. doi: 10.1038/s41587-019-0217-9

White, P. L., Backx, M., and Barnes, R. A. (2017). Diagnosis and management of *Pneumocystis jirovecii* infection. *Expert Rev. Anti-Infect. Ther.* 15, 435–447. doi: 10.1080/14787210.2017.1305887

Wickes, B. L., and Wiederhold, N. P. (2018). Molecular diagnostics in medical mycology. *Nat. Commun.* 9:5135. doi: 10.1038/s41467-018-07556-5

Winand, R., Bogaerts, B., Hoffman, S., Lefevre, L., Delvoye, M., Braekel, J. V., et al. (2019). Targeting the 16s rRNA gene for bacterial identification in complex mixed samples: comparative evaluation of second (Illumina) and third (Oxford Nanopore technologies) generation sequencing technologies. *Int. J. Mol. Sci.* 21, 1–22. doi: 10.3390/ijms21010298

Xiao, C. L., Chen, Y., Xie, S. Q., Chen, K. N., Wang, Y., Han, Y., et al. (2017). MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nat. Methods* 14, 1072–1074. doi: 10.1038/nmeth.4432

Xu, Y., Lewandowski, K., Jeffery, K., Downs, L. O., Foster, D., Sanderson, N. D., et al. (2020). Nanopore metagenomic sequencing to investigate nosocomial transmission of human metapneumovirus from a unique genetic group among haematology patients in the United Kingdom. *J. Inf. Secur.* 80, 571–577. doi: 10.1016/j.jinf.2020.02.003

Yamashita, T., Takeda, H., Takai, A., Arasawa, S., Nakamura, F., Mashimo, Y., et al. (2020). Single-molecular real-time deep sequencing reveals the dynamics of multi-drug resistant haplotypes and structural variations in the hepatitis C virus genome. *Sci. Rep.* 10:2651. doi: 10.1038/s41598-020-59397-2

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Hoang, Irinyi, Hu, Schwessinger and Meyer. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

1.5 Hypothesis

The application of long-read Nanopore sequencing technologies to the identification of pathogenic fungal species will improve mycological diagnostic methods, leading to a drastic reduction in diagnostic turnaround time, enabling timely/tailored antifungal treatment initiation, improve patient outcomes, and greatly reduce associated health care costs.

1.6 Aims

The overall aim of this thesis is to contribute to the improvement of metabarcoding and metagenomic long-read sequencing of pathogenic fungal species.

The specific aims are:

Aim 1: To improve established databases containing reference fungal barcode sequences.

Aim 2: To design a fungal-specific bioinformatic tool for long-read metabarcoding studies.

Aim 3: To investigate the current ability of ONT sequencing technology to identify causative fungal pathogen species in clinical samples.

Aim 4: To optimize DNA extraction and human DNA depletion methods for spiked fungal blood samples to improve sequencing efficiency.

Aim 5: To develop a fungal-specific bioinformatic pipeline for metagenomic studies.

1.7 References

1. Fisher, M.C., *et al.*, *Threats posed by the fungal kingdom to humans, wildlife, and agriculture.* mBio, 2020. **11**(3).
2. Mora, C., *et al.*, *How many species are there on Earth and in the ocean?* PLoS Biol., 2011. **9**(8): p. e1001127.
3. Connell, L., *et al.*, *Fungal diversity associated with an active deep sea volcano: Vailulu'u Seamount, Samoa.* Geomicrobiol. J., 2008. **26**(8): p. 597-605.
4. Robinson, C.H., *Cold adaptation in Arctic and Antarctic fungi.* New Phytol., 2001. **151**(2): p. 341-353.
5. Dighton, J., T. Tugay, and N. Zhdanova, *Fungi and ionizing radiation from radionuclides.* FEMS Microbiol. Lett., 2008. **281**(2): p. 109-20.
6. Onofri, S., *et al.*, *Survival of Antarctic cryptoendolithic fungi in simulated martian conditions on board the International Space Station.* Astrobiology, 2015. **15**(12): p. 1052-9.
7. Peay, K.G., P.G. Kennedy, and J.M. Talbot, *Dimensions of biodiversity in the Earth mycobiome.* Nat. Rev. Microbiol., 2016. **14**(7): p. 434-47.
8. Fisher, M.C., *et al.*, *Emerging fungal threats to animal, plant and ecosystem health.* Nature, 2012. **484**(7393): p. 186-94.
9. Cruz, C.D. and B. Valent, *Wheat blast disease: danger on the move.* Trop. Plant Pathol., 2017. **42**: p. 210-222.
10. Fernandez, J. and K. Orth, *Rise of a cereal killer: The biology of Magnaporthe oryzae biotrophic growth.* Trends Microbiol., 2018. **26**(7): p. 582-597.
11. Ceresini, P.C., *et al.*, *Wheat blast: from its origins in South America to its emergence as a global threat.* Mol. Plant Pathol., 2019. **20**: p. 155-172.

12. Saleh, D., *et al.*, *South-East Asia is the center of origin, diversity and dispersion of the rice blast fungus, Magnaporthe oryzae.* New Phytol., 2013. **201**: p. 1440-1456.

13. Skamnioti, P. and S.J. Gurr, *Against the grain: safeguarding rice from rice blast disease.* Trends Biotechnol., 2009. **27**(3): p. 141-50.

14. Williamson, B., *et al.*, *Botrytis cinerea: the cause of grey mould disease.* Mol. Plant Pathol., 2007. **8**(5): p. 561-80.

15. Legard, D.E., *et al.*, *Effects of plant spacing and cultivar on incidence of Botrytis fruit rot in annual strawberry.* Plant Dis., 2000. **84**(5): p. 531-538.

16. Stuart, S.N., *et al.*, *Status and trends of amphibian declines and extinctions worldwide.* Science, 2004. **306**(5702): p. 1783-6.

17. Fischer, M. and T. Garner, *The relationship between the emergence of Batrachochytrium dendrobatidis, the international trade in amphibians and introduced amphibian species.* Fungal Biol. Rev., 2007. **21**(1): p. 2-9.

18. Fisher, M.C., T.W. Garner, and S.F. Walker, *Global emergence of Batrachochytrium dendrobatidis and amphibian chytridiomycosis in space, time, and host.* Annu. Rev. Microbiol., 2009. **63**: p. 291-310.

19. Geiger, C.C., *et al.*, *Antifungal treatment of wild amphibian populations caused a transient reduction in the prevalence of the fungal pathogen, Batrachochytrium dendrobatidis.* Sci. Rep., 2017. **7**(1): p. 5956.

20. Voyles, J., *et al.*, *Temperature alters reproductive life history patterns in Batrachochytrium dendrobatidis, a lethal pathogen associated with the global loss of amphibians.* Ecol. Evol., 2012. **2**(9): p. 2241-9.

21. Bosch, J., *et al.*, *Successful elimination of a lethal wildlife infectious disease in nature.* Biol. Lett., 2015. **11**(11).

22. Blehert, D.S., *et al.*, *Bat white-nose syndrome: an emerging fungal pathogen?* Science, 2009. **323**(5911): p. 227.

23. Hoyt, J.R., A.M. Kilpatrick, and K.E. Langwig, *Ecology and impacts of white-nose syndrome on bats.* Nat. Rev. Microbiol., 2021. **19**(3): p. 196-210.

24. Brown, G.D., *et al.*, *Hidden killers: human fungal infections.* Sci. Transl. Med., 2012. **4**(165): p. 165rv13.

25. Rodrigues, M.L. and P.C. Albuquerque, *Searching for a change: The need for increased support for public health and research on fungal diseases.* PLoS Negl. Trop. Dis., 2018. **12**(6): p. e0006479.

26. Rammaert, B., *et al.*, *Diabetes and mucormycosis: a complex interplay.* Diabetes Metab., 2012. **38**(3): p. 193-204.

27. Enoch, D.A., *et al.*, *The changing epidemiology of invasive fungal infections.* Methods Mol. Biol., 2017. **1508**: p. 17-65.

28. Suleyman, G. and G.J. Alangaden, *Nosocomial fungal infections: epidemiology, infection control, and prevention.* Infect. Dis. Clin North Am., 2016. **30**(4): p. 1023-1052.

29. Kwon-Chung, K.J., *et al.*, *Cryptococcus neoformans and Cryptococcus gattii, the etiologic agents of cryptococcosis.* Cold Spring Harb. Perspect. Med., 2014. **4**(7): p. a019760.

30. Prevention, C.f.D.C.a., *1993 Revised classification system for HIV infection and expanded surveillance case definition for AIDS among adolescents and adults.* MMWR Recomm. Rep., 1992. **41**: p. 1-19.

31. Chen, S.C.-A., W. Meyer, and T.C. Sorrell, *Cryptococcus gattii infections.* Clin. Microbiol. Rev., 2014. **27**(4): p. 980-1024.

32. Kordossis, T., *et al.*, *First report of Cryptococcus laurentii meningitis and a fatal case of Cryptococcus albidus cryptococcaemia in AIDS patients.* Med. Mycol., 1998. **36**(5): p. 335-9.

33. Edman, J.C., *et al.*, *Ribosomal RNA sequence shows Pneumocystis carinii to be a member of the fungi.* Nature, 1988. **334**(6182): p. 519-22.

34. Sokulska, M., *et al.*, *Pneumocystis jirovecii - from a commensal to pathogen: clinical and diagnostic review.* Parasitol. Res., 2015. **114**(10): p. 3577-85.

35. Phair, J., *et al.*, *The risk of Pneumocystis carinii pneumonia among men infected with human immunodeficiency virus type 1. Multicenter AIDS Cohort Study Group.* N. Engl. J. Med., 1990. **322**(3): p. 161-5.

36. Morris, A., *et al.*, *Current epidemiology of Pneumocystis pneumonia.* Emerg. Infect. Dis., 2004. **10**(10): p. 1713-20.

37. Ghembaza, A., *et al.*, *Risk factors and prevention of Pneumocystis jirovecii pneumonia in patients with autoimmune and inflammatory diseases.* Chest, 2020. **158**(6): p. 2323-2332.

38. Salzer, H.J.F., *et al.*, *Clinical, diagnostic, and treatment disparities between HIV-infected and non-HIV-infected immunocompromised patients with Pneumocystis jirovecii pneumonia.* Respiration, 2018. **96**(1): p. 52-65.

39. Roblot, F., *et al.*, *Analysis of underlying diseases and prognosis factors associated with Pneumocystis carinii pneumonia in immunocompromised HIV-negative patients.* Eur. J. Clin. Microbiol. Infect. Dis., 2002. **21**(7): p. 523-31.

40. Green, H., *et al.*, *Prophylaxis of Pneumocystis pneumonia in immunocompromised non-HIV-infected patients: systematic review and meta-analysis of randomized controlled trials.* Mayo Clin. Proc., 2007. **82**(9): p. 1052-9.

41. Martins, N., *et al.*, *Candidiasis: predisposing factors, prevention, diagnosis and alternative treatment.* Mycopathologia, 2014. **177**(5-6): p. 223-40.

42. Wisplinghoff, H., *et al.*, *Nosocomial bloodstream infections in US hospitals: analysis of 24,179 cases from a prospective nationwide surveillance study.* Clin. Infect. Dis., 2004. **39**(3): p. 309-17.

43. Pappas, P.G., *et al.*, *Clinical practice guideline for the management of Candidiasis: 2016 update by the infectious diseases society of America.* Clin. Infect. Dis., 2016. **62**(4): p. e1-50.

44. Vincent, J.L., *et al.*, *International study of the prevalence and outcomes of infection in intensive care units.* JAMA, 2009. **302**(21): p. 2323-9.

45. Logan, C., I. Martin-Loeches, and T. Bicanic, *Invasive candidiasis in critical care: challenges and future directions.* Intensive Care Med., 2020. **46**(11): p. 2001-2014.

46. Strollo, S., *et al.*, *Epidemiology of hospitalizations associated with invasive candidiasis, United States, 2002–2012.* Emerg. Infect. Dis., 2017. **23**(1): p. 7-13.

47. Castanheira, M., *et al.*, *Antifungal susceptibility patterns of a global collection of fungal isolates: results of the SENTRY Antifungal Surveillance Program (2013).* Diagn Microbiol Infect Dis, 2016. **85**(2): p. 200-4.

48. Pfaller, M.A., *et al.*, *Geographic variations in species distribution and echinocandin and azole antifungal resistance rates among Candida bloodstream infection isolates: report from the SENTRY Antimicrobial Surveillance Program (2008 to 2009).* J. Clin. Microbiol., 2011. **49**(1): p. 396-9.

49. Lamoth, F., *et al.*, *Changes in the epidemiological landscape of invasive candidiasis.* J. Antimicrob. Chemother., 2018. **73**(suppl_1): p. i4-i13.

50. Pappas, P.G., *et al.*, *Invasive candidiasis.* Nat. Rev. Dis. Primers, 2018. **4**: p. 18026.

51. Du, H., *et al.*, *Candida auris: Epidemiology, biology, antifungal resistance, and virulence.* PLoS Pathog., 2020. **16**(10): p. e1008921.

52. Spivak, E.S. and K.E. Hanson, *Candida auris: an emerging fungal pathogen.* J. Clin. Microbiol., 2018. **56**(2).

53. Cadena, J., G.R. Thompson, 3rd, and T.F. Patterson, *Invasive aspergillosis: current strategies for diagnosis and management.* Infect. Dis. Clin. North Am., 2016. **30**(1): p. 125-42.

54. Denis, B., *et al.*, *Relevance of EORTC criteria for the diagnosis of invasive aspergillosis in HIV-infected patients, and survival trends over a 20-year period in France.* Clin. Infect. Dis., 2015. **61**(8): p. 1273-80.

55. Mohammed, A.P., *et al.*, *An unusual case of invasive aspergillosis in an immunocompetent individual.* BMJ Case Rep., 2015. **2015**.

56. Cheon, S., *et al.*, *Disseminated aspergillosis in the immunocompetent host: a case report and literature review.* Mycopathologia, 2015. **180**(3-4): p. 217-22.

57. Tudesq, J.J., *et al.*, *Invasive pulmonary aspergillosis in nonimmunocompromised hosts.* Semin. Respir. Crit. Care Med., 2019. **40**(4): p. 540-547.

58. Neofytos, D., *et al.*, *Epidemiology and outcome of invasive fungal infection in adult hematopoietic stem cell transplant recipients: analysis of Multicenter Prospective Antifungal Therapy (PATH) Alliance registry.* Clin. Infect. Dis., 2009. **48**(3): p. 265-73.

59. Patterson, J.E., *et al.*, *Investigation and control of aspergillosis and other filamentous fungal infections in solid organ transplant recipients.* Transpl. Infect. Dis., 2000. **2**(1): p. 22-8.

60. Bulpa, P., *et al.*, *Invasive pulmonary aspergillosis in chronic obstructive pulmonary disease exacerbations.* Semin. Respir. Crit. Care Med., 2020. **41**(6): p. 851-861.

61. Ader, F., *et al.*, *Invasive pulmonary aspergillosis in chronic obstructive pulmonary disease: an emerging fungal pathogen.* Clin. Microbiol. Infect., 2005. **11**(6): p. 427-9.

62. Dimopoulos, G., *et al.*, *Invasive aspergillosis in the intensive care unit.* Ann. N. Y. Acad. Sci., 2012. **1272**: p. 31-9.

63. Meis, J.F., *et al.*, *Clinical implications of globally emerging azole resistance in Aspergillus fumigatus.* Philos. Trans. R. Soc. Lond. B. Biol. Sci., 2016. **371**(1709).

64. Wiederhold, N.P. and T.F. Patterson, *Emergence of azole resistance in Aspergillus.* Semin. Respir. Crit. Care Med., 2015. **36**(5): p. 673-80.

65. Petrikkos, G., *et al.*, *Epidemiology and clinical manifestations of mucormycosis.* Clin. Infect. Dis., 2012. **54 Suppl 1**: p. S23-34.

66. Jeong, W., *et al.*, *The epidemiology and clinical manifestations of mucormycosis: a systematic review and meta-analysis of case reports.* Clin. Microbiol. Infect., 2019. **25**(1): p. 26-34.

67. Patel, A., *et al.*, *Multicenter epidemiologic study of Coronavirus disease-associated mucormycosis, India.* Emerg. Infect. Dis., 2021. **27**(9): p. 2349-2359.

68. Deutsch, P.G., J. Whittaker, and S. Prasad, *Invasive and non-invasive fungal rhinosinusitis-a review and update of the evidence.* Medicina (Kaunas), 2019. **55**(7).

69. Petrikkos, G., *et al.*, *Mucormycosis: ten-year experience at a tertiary-care center in Greece.* Eur. J. Clin. Microbiol. Infect. Dis., 2003. **22**(12): p. 753-6.

70. Rammaert, B., *et al.*, *Healthcare-associated mucormycosis.* Clin. Infect. Dis., 2012. **54 Suppl 1**: p. S44-54.

71. Neblett Fanfair, R., *et al.*, *Necrotizing cutaneous mucormycosis after a tornado in Joplin, Missouri, in 2011.* N. Engl. J. Med., 2012. **367**(23): p. 2214-25.

72. Slavin, M., *et al.*, *Invasive infections due to filamentous fungi other than Aspergillus: epidemiology and determinants of mortality.* Clin. Microbiol. Infect., 2015. **21**(5): p. 490 e1-10.

73. Roden, M.M., *et al.*, *Epidemiology and outcome of zygomycosis: a review of 929 reported cases.* Clin. Infect. Dis., 2005. **41**(5): p. 634-53.

74. Chakrabarti, A., *et al.*, *The rising trend of invasive zygomycosis in patients with uncontrolled diabetes mellitus.* Med. Mycol., 2006. **44**(4): p. 335-42.

75. Singh, A.K., *et al.*, *Mucormycosis in COVID-19: A systematic review of cases reported worldwide and in India.* Diabetes Metab. Syndr., 2021. **15**(4): p. 102146.

76. Leibovitch, E.C. and S. Jacobson, *Vaccinations for neuroinfectious disease: A global health priority.* Neurotherapeutics, 2016. **13**(3): p. 562-70.

77. Nami, S., *et al.*, *Fungal vaccines, mechanism of actions and immunology: A comprehensive review.* Biomed. Pharmacother., 2019. **109**: p. 333-344.

78. Hamad, M., *Universal fungal vaccines: could there be light at the end of the tunnel?* Hum. Vaccin. Immunother., 2012. **8**(12): p. 1758-63.

79. Medici, N.P. and M. Del Poeta, *New insights on the development of fungal vaccines: from immunity to recent challenges.* Mem. Inst. Oswaldo Cruz, 2015. **110**(8): p. 966-73.

80. Mellinghoff, S.C., *et al.*, *Primary prophylaxis of invasive fungal infections in patients with haematological malignancies: 2017 update of the recommendations of the Infectious Diseases Working Party (AGIHO) of the German Society for Haematology and Medical Oncology (DGHO).* Ann. Hematol., 2018. **97**(2): p. 197-207.

81. Lehrnbecher, T., *et al.*, *Extended dosing regimens for fungal prophylaxis.* Clin. Microbiol. Rev., 2019. **32**(3).

82. Worth, L.J., *et al.*, *Optimizing antifungal drug dosing and monitoring to avoid toxicity and improve outcomes in patients with haematological disorders.* Intern. Med. J., 2008. **38**(6b): p. 521-37.

83. Maertens, J., *Evaluating prophylaxis of invasive fungal infections in patients with haematologic malignancies.* Eur. J. Haematol., 2007. **78**(4): p. 275-82.

84. Kontoyiannis, D.P., *et al.*, *Zygomycosis in a tertiary-care cancer center in the era of Aspergillus-active antifungal therapy: a case-control observational study of 27 recent cases.* J. Infect. Dis., 2005. **191**(8): p. 1350-60.

85. Ashbee, H.R., *et al.*, *Therapeutic drug monitoring (TDM) of antifungal agents: guidelines from the British Society for Medical Mycology.* J. Antimicrob. Chemother., 2014. **69**(5): p. 1162-76.

86. Fisher, M.C., *et al.*, *Worldwide emergence of resistance to antifungal drugs challenges human health and food security.* Science, 2018. **360**(6390): p. 739-742.

87. Revie, N.M., *et al.*, *Antifungal drug resistance: evolution, mechanisms and impact.* Curr. Opin. Microbiol., 2018. **45**: p. 70-76.

88. Schelenz, S., *et al.*, *National mycology laboratory diagnostic capacity for invasive fungal diseases in 2017: Evidence of sub-optimal practice.* J. Infect., 2019. **79**(2): p. 167-173.

89. Barnes, R.A., *Early diagnosis of fungal infection in immunocompromised patients.* J. Antimicrob. Chemother., 2008. **61 Suppl 1**: p. i3-6.

90. von Eiff, M., *et al.*, *Pulmonary aspergillosis: early diagnosis improves survival.* Respiration, 1995. **62**(6): p. 341-7.

91. Austin, B., *The value of cultures to modern microbiology.* Antonie Van Leeuwenhoek, 2017. **110**(10): p. 1247-1256.

92. Kurtzman, C., J.W. Fell, and T. Boekhout, *The yeasts: a taxonomic study*. 2011, Amsterdam, Netherlands: Elsevier.

93. Guarner, J. and M.E. Brandt, *Histopathologic diagnosis of fungal infections in the 21st century.* Clin. Microbiol. Rev., 2011. **24**(2): p. 247-80.

94. Schelenz, S., *et al.*, *British Society for Medical Mycology best practice recommendations for the diagnosis of serious fungal diseases.* Lancet Infect. Dis., 2015. **15**(4): p. 461-74.

95. Irinyi, L., *et al.*, *DNA barcoding of fungi causing infections in humans and animals.* Fungal Biol., 2016. **120**(2): p. 125-36.

96. Dominguez, S.R., *et al.*, *Comparison of Herpes Simplex Virus PCR with culture for virus detection in multisource surface swab specimens from neonates.* J. Clin. Microbiol., 2018. **56**(10).

97. Gustafson, E., *et al.*, *DNA-based detection for onychomycosis correlates better to histopathology than does fungal culture.* Dermatol. Online J., 2019. **25**(7): p. 1-9.

98. Kung, V., R. Chernock, and C.-A. Burnham, *Diagnostic accuracy of fungal identification in histopathology and cytopathology specimens.* Eur. J. Clin. Microbiol. Infect. Dis., 2018. **37**(1): p. 157-165.

99. Goldstein, E.J., K.L. Tyrrell, and D.M. Citron, *Lactobacillus species: taxonomic complexity and controversial susceptibilities.* Clin. Infect. Dis., 2015. **60 Suppl 2**: p. S98-107.

100. De Hoog, G., *et al.*, *Atlas of Clinical Fungi 2019*. 2019: Reus: Utrecht, The Netherlands.

101. Li, L., *et al.*, *The importance of the viable but non-culturable state in human bacterial pathogens.* Front. Microbiol., 2014. **5**: p. 258.

102. White, P.L., M. Backx, and R.A. Barnes, *Diagnosis and management of Pneumocystis jirovecii infection.* Expert Rev. Anti. Infect. Ther., 2017. **15**(5): p. 435-447.

103. Murray, I.G., *Some aspects of the biochemical differentiation of pathogenic fungi: A review.* J. Gen. Microbiol., 1968. **52**: p. 213-221.

104. Taylor, J.B., *Biochemical tests for identification of mycelial cultures of basidiomycetes.* Ann. Appl. Biol., 1974. **78**: p. 113-123.

105. Grant, M.L., *et al.*, *Comparative evaluation of the BD Phoenix Yeast ID Panel and Remel RapID Yeast Plus System for yeast identification.* Can. J. Infect. Dis. Med. Microbiol., 2016. **2016**: p. 4094932.

106. Pincus, D.H., S. Orenga, and S. Chatellier, *Yeast identification--past, present, and future methods.* Med. Mycol., 2007. **45**(2): p. 97-121.

107. Posteraro, B., *et al.*, *Are the conventional commercial yeast identification methods still helpful in the era of new clinical microbiology diagnostics? A meta-analysis of their accuracy.* J. Clin. Microbiol., 2015. **53**(8): p. 2439-50.

108. Melhem, M.S., *et al.*, *Use of the VITEK 2 system to identify and test the antifungal susceptibility of clinically relevant yeast species.* Braz. J. Microbiol., 2013. **44**(4): p. 1257-1266.

109. Arastehfar, A., *et al.*, *Comparison of 21-Plex PCR and API 20C AUX, MALDI-TOF MS, and rDNA sequencing for a wide range of clinically isolated yeast species: Improved identification by combining 21-Plex PCR and API 20C AUX as an alternative strategy for developing countries.* Front. Cell. Infect. Microbiol., 2019. **9**: p. 21.

110. Wanger, A., *et al.*, *Chapter 5-Biochemical tests and staining techniques for microbial identification.* Microbiology and Molecular Diagnosis in Pathology. Elsevier, 2017.

111. Vidal, J.E. and D.R. Boulware, *Lateral flow assay for cryptococcal antigen: An important advance to improve the continuum of HIV care and reduce cryptococcal meningitis-related mortality.* Rev. Inst. Med. Trop. Sao Paulo, 2015. **57 Suppl 19**: p. 38-45.

112. Kozel, T.R. and B. Wickes, *Fungal diagnostics.* Cold Spring Harb. Perspect. Med., 2014. **4**(4): p. a019299.

113. Richardson, M. and I. Page, *Role of serological tests in the diagnosis of mold infections.* Curr. Fungal Infect. Rep., 2018. **12**(3): p. 127-136.

114. Fenselau, C. and P.A. Demirev, *Characterization of intact microorganisms by MALDI mass spectrometry.* Mass Spectrom. Rev., 2001. **20**(4): p. 157-71.

115. Clark, A.E., *et al.*, *Matrix-assisted laser desorption ionization-time of flight mass spectrometry: a fundamental shift in the routine practice of clinical microbiology.* Clin. Microbiol. Rev., 2013. **26**(3): p. 547-603.

116. Angeletti, S., *Matrix assisted laser desorption time of flight mass spectrometry (MALDI-TOF MS) in clinical microbiology.* J. Microbiol. Methods., 2017. **138**: p. 20-29.

117. Sauget, M., *et al.*, *Can MALDI-TOF mass spectrometry reasonably type bacteria?* Trends Microbiol., 2017. **25**(6): p. 447-455.

118. Patel, T.S., *et al.*, *Cost analysis of implementing matrix-assisted laser desorption ionization-time of flight mass spectrometry plus real-time antimicrobial stewardship intervention for bloodstream infections.* J. Clin. Microbiol., 2017. **55**(1): p. 60-67.

119. Sow, D., *et al.*, *Usefulness of MALDI-TOF mass spectrometry for routine identification of Candida species in a resource-poor setting.* Mycopathologia, 2015. **180**(3-4): p. 173-9.

120. Stein, M., *et al.*, *Evaluation of three MALDI-TOF mass spectrometry libraries for the identification of filamentous fungi in three clinical microbiology laboratories in Manitoba, Canada.* Mycoses, 2018. **61**(10): p. 743-753.

121. Luethy, P.M. and A.M. Zelazny, *Rapid one-step extraction method for the identification of molds using MALDI-TOF MS.* Diagn. Microbiol. Infect. Dis., 2008. **91**(2): p. 130-135.

122. Dendis, M., *et al.*, *PCR-RFLP detection and species identification of fungal pathogens in patients with febrile neutropenia.* Clin. Microbiol. Infect., 2003. **9**(12): p. 1191-202.

123. Marcos-Arias, C., *et al.*, *Utility of two PCR-RFLP-based techniques for identification of Candida parapsilosis complex blood isolates.* Mycoses, 2020. **63**(5): p. 461-470.

124. Li, J.S., *et al.*, *PCR-RFLP for Detection of Fusarium graminearum genotypes with resistance to phenamacril.* Plant Dis., 2021. **105**(4): p. 889-895.

125. Brandt, M.E., *et al.*, *Utility of random amplified polymorphic DNA PCR and TaqMan automated detection in molecular identification of Aspergillus fumigatus.* J. Clin. Microbiol., 1998. **36**(7): p. 2057-62.

126. Yamamoto, Y., *et al.*, *Random amplified polymorphic DNA analysis of clinically and environmentally isolated Cryptococcus neoformans in Nagasaki.* J. Clin. Microbiol., 1995. **33**(12): p. 3328-32.

127. Robert, F., *et al.*, *Use of random amplified polymorphic DNA as a typing method for Candida albicans in epidemiological surveillance of a burn unit.* J. Clin. Microbiol., 1995. **33**(9): p. 2366-71.

128. Lindsley, M.D., *et al.*, *Rapid identification of dimorphic and yeast-like fungal pathogens using specific DNA probes.* J. Clin. Microbiol., 2001. **39**(10): p. 3505-11.

129. Sandhu, G.S., *et al.*, *Molecular probes for diagnosis of fungal infections.* J. Clin. Microbiol., 1995. **33**(11): p. 2913-9.

130. Lieckfeldt, E., W. Meyer, and T. Borner, *Rapid identification and differentiation of yeasts by DNA and PCR fingerprinting.* J. Basic Microbiol., 1993. **33**(6): p. 413-25.

131. Meyer, W., K. Maszewska, and T.C. Sorrell, *PCR fingerprinting: a convenient molecular tool to distinguish between Candida dubliniensis and Candida albicans.* Med. Mycol., 2001. **39**(2): p. 185-93.

132. Klingspor, L. and S. Jalal, *Molecular detection and identification of Candida and Aspergillus spp. from clinical samples using real-time PCR.* Clin. Microbiol. Infect., 2006. **12**(8): p. 745-753.

133. Mai, M., *et al.*, *Real-time PCR-based identification of bacterial and fungal pathogens from blood samples.* Methods Mol. Biol., 2015. **1237**: p. 139-47.

134. Neely, L.A., *et al.*, *T2 magnetic resonance enables nanoparticle-mediated rapid detection of candidemia in whole blood.* Sci. Transl. Med., 2013. **5**(182): p. 182ra54.

135. Mylonakis, E., *et al.*, *T2 magnetic resonance assay for the rapid diagnosis of candidemia in whole blood: a clinical trial.* Clin. Infect. Dis., 2015. **60**(6): p. 892-9.

136. Clancy, C.J., *et al.*, *Detecting infections rapidly and easily for candidemia trial, part 2 (DIRECT2): A prospective, multicenter study of the T2Candida Panel.* Clin. Infect. Dis., 2018. **66**(11): p. 1678-1686.

137. Hebert, P.D., *et al.*, *Biological identifications through DNA barcodes.* Proc. Biol. Sci., 2003. **270**(1512): p. 313-321.

138. Kimura, M., *A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences.* J. Mol. Evol., 1980. **16**(2): p. 111-120.

139. Kurtzman, C.P. and C.J. Robnett, *Identification and phylogeny of ascomycetous yeasts from analysis of nuclear large subunit (26S) ribosomal DNA partial sequences.* Antonie Van Leeuwenhoek, 1998. **73**(4): p. 331-71.

140. Fell, J.W., *et al.*, *Biodiversity and systematics of basidiomycetous yeasts as determined by large-subunit rDNA D1/D2 domain sequence analysis.* Int. J. Syst. Evol. Microbiol., 2000. **50 Pt 3**: p. 1351-1371.

141. Scorzetti, G., *et al.*, *Systematics of basidiomycetous yeasts: a comparison of large subunit D1/D2 and internal transcribed spacer rDNA regions.* FEMS Yeast Res., 2002. **2**(4): p. 495-517.

142. Dettman, J.R., F.M. Harbinski, and J.W. Taylor, *Ascospore morphology is a poor predictor of the phylogenetic relationships of Neurospora and Gelasinospora.* Fungal Genet. Biol., 2001. **34**(1): p. 49-61.

143. Xu, J., *Fungal DNA barcoding.* Genome, 2016. **59**(11): p. 913-932.

144. O'Donnell, K., *et al.*, *Genealogical concordance between the mating type locus and seven other nuclear genes supports formal recognition of nine phylogenetically distinct species within the Fusarium graminearum clade.* Fungal Genet. Biol., 2004. **41**(6): p. 600-23.

145. Short, D.P., *et al.*, *Phylogenetic relationships among members of the Fusarium solani species complex in human infections and the descriptions of F. keratoplasticum sp. nov. and F. petroliphilum stat. nov.* Fungal Genet. Biol., 2013. **53**: p. 59-70.

146. Schoch, C.L., *et al.*, *Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi.* Proc. Natl. Acad. Sci. U. S. A., 2012. **109**(16): p. 6241-6246.

147. White, T., J., *et al.*, *Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics*, in *PCR Protocols: A guide to methods and applications*. 1990, Academic Press: New York. p. 315-322.

148. Group, C.P.W., *A DNA barcode for land plants.* Proc. Natl. Acad. Sci. U. S. A., 2009. **106**(31): p. 12794-7.

149. Vilgalys, R. and M. Hester, *Rapid genetic identification and mapping of enzymatically amplified ribosomal DNA from several Cryptococcus species.* J. Bacteriol., 1990. **172**(8): p. 4238-46.

150. Stielow, J.B., *et al.*, *One fungus, which genes? Development and assessment of universal primers for potential secondary fungal DNA barcodes.* Persoonia, 2015. **35**: p. 242.

151. Irinyi, L., *et al.*, *International Society of Human and Animal Mycology (ISHAM)-ITS reference DNA barcoding database—the quality controlled standard tool for routine identification of human and animal pathogenic fungi.* Med. Mycol., 2015. **53**(4): p. 313-337.

152. Eberhardt, U., *A constructive step towards selecting a DNA barcode for fungi.* New Phytol., 2010. **187**(2): p. 265-268.

153. Balajee, S.A., *et al.*, *Sequence-based identification of Aspergillus, Fusarium, and Mucorales species in the clinical mycology laboratory: where are we and where should we go from here?* J. Clin. Microbiol., 2009. **47**(4): p. 877-84.

154. Heinrichs, G., G.S. de Hoog, and G. Haase, *Barcode identifiers as a practical tool for reliable species assignment of medically important black yeast species.* J. Clin. Microbiol., 2012. **50**(9): p. 3023-30.

155. Lewis, C., *et al.*, *Identification of fungal DNA barcode targets and PCR primers based on Pfam protein families and taxonomic hierarchy.* The Open Applied Informatics Journal, 2011. **511**.

156. Robert, V., *et al.*, *The quest for a general and reliable fungal DNA barcode.* The Open Applied Informatics Journal, 2011. **511**.

157. Sasikumar, A.N., W.B. Perez, and T.G. Kinzy, *The many roles of the eukaryotic elongation factor 1 complex.* Wiley Interdiscip. Rev. RNA, 2012. **3**(4): p. 543-55.

158. Roger, A.J., *et al.*, *An evaluation of elongation factor 1 alpha as a phylogenetic marker for eukaryotes.* Mol. Biol. Evol., 1999. **16**(2): p. 218-33.

159. Nouripour-Sisakht, S., *et al.*, *Characterization of the translation elongation factor 1-alpha gene in a wide range of pathogenic Aspergillus species.* J. Med. Microbiol., 2017. **66**(4): p. 419-429.

160. Mirhendi, H., *et al.*, *Translation elongation factor 1-α gene as a potential taxonomic and identification marker in dermatophytes.* Med. Mycol., 2015. **53**(3): p. 215-24.

Chapter 2
Reference DNA Barcode Sequence Generation

2.1 Introduction

Implementation of DNA barcoding relies upon established accurate reference barcode databases for comparison to generated sample sequences for accurate identification. There have been several publicly available databases established for the compilation of sequences, however, many do not meet the criteria to be the ideal fungal DNA barcoding database. Such a database would be publicly accessible without cost and provide accurate reference barcode sequences with high taxonomic and geographic coverage alongside relevant metadata associated with the samples [1]. The ideal fungal DNA barcoding database would contain reference sequencings for both the primary and secondary fungal DNA barcode. A curated collection would aid in the flexibility of the database to adjust to any taxonomic changes that may occur. An additional feature to accommodate the taxonomic clustering of unidentified sequences may be beneficial, as the increase in metagenomic studies has resulted in the discovery of sequences unattributed to any species. Although a database may not meet all requirements, the benefits and drawbacks of sequence databases containing fungal barcode sequences are discussed below.

2.1.1 GenBank

Available at *www.ncbi.nlm.nih.gov/genbank/*

GenBank has remained the prominent, comprehensive public database since its conception in 1982 by the National Center for Biotechnology Information (NCBI). Alongside the EMBL-EBI European Nucleotide Archive and the DNA Databank of Japan, to form the International Nucleotide Sequence Database Collaboration (INSDC) [2]. The database contains over 6.25 trillion base pairs amongst 1.6 billion nucleotide sequences, over 1.5 million of which are fungal sequences, and continues to undergo expansive growth with a 69.25% increase from 2019-2020 [3]. Sequences submitted to GenBank undergo quality

checks for contamination, proper translation of coding regions, correct taxonomy, and bibliographic citations [4]. The draft submissions are then sent to the original authors to review and are given an accession number as confirmation of successful submission. Bacterial sequences are given an additional quality control method using average nucleotide identity analysis during submission and as a tool to scan database entries for incorrect taxonomic assignments [3]. Minimal sample data is required for submission and as such there is a growing proportion of sequences that are not identified to the species level. For fungi, this was demonstrated as 35% of fungal ITS region sequences were found to be insufficiently identified in 2009, and this proportion is estimated to have increased due to whole-genome shotgun sequencing [5]. Sequences deposited into GenBank are not rigorously quality controlled for taxonomic accuracy and more than 2,000,000 submissions are likely contaminated resulting in inaccurate species identification when using GenBank as a reference sequence database [6-9]. Additionally, it was previously found that 20% of the ITS entries in the INSDC were incorrectly identified to the species level and a large proportion of entries lack updated annotation and strain information [10].

2.1.2 BOLD

Available at *boldsystems.org*

The Barcode of Life Data System (BOLD) was established by the Consortium for the Barcode of Life to build a barcode library for all eukaryotic life [11]. The large global collaborative effort currently boasts 10 million stored barcodes, primarily spanning animal and plant species. The data system supports four functions: the database for barcodes and associated metadata, an educational portal, a barcode index number database for sequence clusters used to approximate species and a data analysis function to assemble and validate barcode sequences [12]. Additionally, all barcode sequences contained in the BOLD database

are submitted to GenBank for wider reach. Fungal barcodes account for less than 24,000 barcodes in the BOLD database and so a substantial effort would be required to reach the coverage already attained for animal and plant species in the database. Sequences submitted to BOLD are translated into amino acids and automatically checked for stop codons and common contaminants. Those with errors or contaminants may then be manually corrected or reassembled from chromatogram trace files [13]. Minimal metadata is required for submission and sequences do not require associated published literature. Although BOLD contains a large collection of reference barcode sequences, the database is not curated to ensure accurate barcode sequences alongside relevant metadata, giving rise to potential errors in accurate identification.

2.1.3 RefSeq

Available at *https://www.ncbi.nlm.nih.gov/refseq/*

RefSeq is a public database of nucleotide and protein sequences initiated and run by NCBI [14]. It aims to contain an extensive database of accurate and full-length sequence data, accurate sequence to gene correspondence, and to note additional feature annotation regions or events of interest [14]. Although run by the same organisation, RefSeq differs from GenBank by its highly curated nature. The database is a large collaborative effort, as sections of the database are curated and maintained by NCBI staff whilst other sections are run by specialist collaborators to collect sequences and gene and functional information [15]. RefSeq also features in-house annotation tools that are constantly being improved upon [15]. Currently, RefSeq contains 206 million sequences covering 93,000 taxa and has an annual sequence growth of 26% [16]. The fungal sequences contained in the database amount to more than 4.1 million, of which 4 million are mRNA sequences. RefSeq entries are sequences taken from GenBank and the sequences are verified for sequence quality, type

status, taxonomy sequence identification, region length, and protein translation. Additional quality control checks are done for different organisms or data types [15, 17]. As such, RefSeq contains high-quality reference sequences. However, the major drawback is the lack of sequence representation, as generally there is only one sequence representative of a given species generated from the type culture of the species [15].

2.1.4 UNITE

Available at *https://unite.ut.ee/*

The UNITE database was established in 2003 to provide ITS sequences from ectomycorrhizal fungi [18, 19]. The initial goal was to reach a collection of 300,000 quality-controlled sequences mined from GenBank by December 2012 [20]. The guidelines for inclusions for these sequences were extremely strict with the associated fungal cultures must be mature specimens with identifiable characteristics and maintained in public collections. Additionally, submissions were limited to samples from northern Europe, and they must have been identified and submitted by a mycological expert [18]. Since its establishment, the limitations on UNITE submissions have been expanded by no longer solely catering to fungi and supporting all eukaryotic submissions, and lifting the geographical restrictions [21, 22]. Importantly, environmental samples and unidentified samples are now allowed to be submitted and UNITE supports hypothetical species [23]. Operational taxonomic units are regularly clustered into Species Hypothesis which can then be used and referenced unambiguously despite the lack of an official taxonomic name [23]. Currently, the UNITE database contains over 2 million fungal ITS sequences attributed to 120,183 species hypotheses. The sequence number has exponentially grown, having doubled since 2019 [23]. Entries are subjected to quality control measures to reject non-ITS and chimeric sequences

and standards are implemented for sequences and metadata. The new state of UNITE has drawn comparisons to become the ITS equivalent of the GenBank database [17].

2.1.5 ISHAM-ITS Database

Available at *https://its.mycologylab.org*

The ISHAM-ITS Database is the result of an international effort by the ISHAM barcoding of pathogenic fungi working group to establish a reliable database of ITS sequences from human and animal pathogenic fungal species [1]. The ISHAM-ITS Database allows users to query an unknown fungal sequence and generates potential matches based on sequence homology searches through the use of a modified BLAST algorithm [24]. The database also provides a confidence score of up to five stars to each query result based on probability value, similarity percentage, fragment number, overlap percentage, and sequence direction, allowing users to easily access the value of the results. Additional strain information such as source, country of origin, and other associated collection numbers are available and searchable.

Upon inception, the ISHAM-ITS database contained 2800 full-length ITS reference sequences representing 421 fungal species. Of these, 245 species were represented by multiple strains, and 176 species were only represented by one strain [1]. Importantly, reference sequences from the type strains of 71 fungal species were deposited to the ISHAM-ITS Database and submitted to RefSeq. The initial study also highlighted fungal species where the ITS region alone was not adequate in providing accurate identification such as *Clavispora lusitaniae, Kodamaea ohmeri,* members of the *Fusarium solani* species complex, and *Scedosporium* species.

The ISHAM-ITS Database is a foundation of reference sequences required for the reliable use of DNA barcoding for the identification of pathogenic fungal species, due to its highly curated and quality-controlled nature. However, improvements are still required to increase

the accuracy and practicality of the database. The ITS reference sequences within the database do not represent all pathogenic fungal species and so efforts to provide reference barcode sequences for those species are required. Additionally, a large proportion of fungal species are represented by one strain and so more reference sequences belonging to those species would benefit the database. Most importantly, the establishment of the secondary fungal DNA barcode, the *TEF1α* region is not represented within the ISHAM-ITS Database and must be updated to have quality-controlled reference secondary barcode sequences deposited into the database, adequately representing all pathogenic fungal species.

2.2 Methods

Primary and secondary fungal DNA barcodes were generated from fungal strains within the Westmead Medical Mycology Collection (WM), the University of Sydney WFCC registration number: WM-1205. This work was partially performed as part of my Honours Thesis "Application of next-generation sequencing to DNA barcoding of pathogenic fungi", submitted in 2017, and expended herein.

2.2.1 Fungal Culturing

Fungal species in the WM Culture Collection were previously identified by traditional identification techniques, such as microscopy. Water cultures were subcultured by Krystyna Maszewska onto Saboraud agar. Filamentous fungi were incubated at 27°C for 2-3 days or until sufficient growth of hyphae was achieved with minimal sporulation. Yeasts were incubated at 37°C for 1-3 days. Cultures were visually checked for purity and if contaminated subculturing from water cultures was repeated.

2.2.2 DNA Extraction

A modified DNA extraction protocol was used based on the phenol-chloroform extract protocol published by Ferrer *et al.* [25].

Two loops of pure fungal culture were placed into a 1.5 mL microfuge tube. To disrupt fungal cell walls, yeast cultures were incubated at -20°C overnight and filamentous fungal material was frozen with liquid nitrogen and ground using a plastic pestle to a fine powder texture. Fungal cultures were then thawed at room temperature and 500 µL lysis buffer (0.5% w/v sodium dodecyl sulphate, 1.2 mM EDTA, 20% v/v 1M Tris-HCl) and 5 µL 2-mercaptoethanol was added. The solution was then mixed thoroughly by vortex for 1 minute and incubated at 65°C for 1 hour. 500 µL of phenol-chloroform-isoamyl alcohol (25:24:1, v/v/v) was added and mixed by vortex for 2 minutes or until a homogenous solution was reached. The solution was then centrifuged for 15 minutes at 14,000 rpm. The aqueous layer was pipetted into a new 1.5 mL microfuge tube, and 500 µL of phenol:chloroform:isoamyl alcohol (25:24:1, v/v/v) was added and mixed to form a homogenous solution. The solution was again centrifuged for 15 minutes at 14,000 rpm. The aqueous layer was pipetted into a new 1.5 mL microfuge tube and mixed with an equal volume of isopropanol and incubated at -20°C overnight. The solution was then centrifuged for 15 minutes at 4°C at 14,000 rpm. The supernatant was removed, and the pellet was washed with 70% ethanol solution twice. The solution was centrifuged for 15 minutes at 14,000 rpm and the supernatant was removed. The pellet was dried for 8-10 minutes at 45°C or until the remaining ethanol had evaporated. DNA was then resuspended in 100 µL of sterile water at room temperature and stored at 4°C for short-term storage and -20°C for long-term storage. The concentration and purity of the resulting DNA were measured by spectrophotometry on the DeNovix DS-11+ Spectrophotometer.

2.2.3 ITS Barcode Amplification

The primary fungal DNA barcode, the ITS region, was amplified from extracted DNA using PCR. All PCR reactions were composed of 5 µL 10X Reaction Buffer, 5 µL dNTP mix, 4 µL 10ng/µL forward and reverse primers, 3 µL 50 mM $MgCl_2$, 15 µL 10ng/µL genomic DNA, 0.5 µL Taq polymerase and nuclease-free water to a total volume of 50 µL. Two primer pairs and the corresponding PCR conditions were used to amplify similar regions (Chapter 1 Figure 2).

The primer pair SR6R (5' AAGTATAAGTCGTAACAAGG '3) and LR1 ((5' GGTTGGTTTCTTTCCT '3) were used under the PCR conditions as follows: 3 minutes initial denaturing at 97°C, followed by 30 cycles of 55 seconds at 97°C, 45 seconds minute annealing at 50°C, 45 seconds at 72°C and final extension at 72°C for 7 minutes [26].

The primer pairs ITS1 (5' TCCGTAGGTGAACCTGCGG' 3) and ITS4 (5' TCCTCCGCTTATTGATATGC' 3) were also used to amplify the ITS region under the PCR conditions as follows: 3 minutes initial denaturing at 97°C, followed by 35 cycles of 45 seconds at 95°C, 45 seconds annealing at 55°C and 45 seconds at 72°C with final extension for 8 minutes at 72°C [27]. All PCR reactions were performed in a Sensoquest thermocycler.

2.2.4 *TEF1α* Barcode Amplification

PCR was also used to amplify the secondary fungal DNA barcode, the *TEF1α* region. All PCR reactions were performed in a total volume of 50 µL, composed of 5 µL 10 X Reaction Buffer, 5 µL dNTP mix, 4 µL 10 ng/µL forward and reverse primers, 3 µL 50 mM $MgCl_2$, 15 µL 10 ng/µL genomic DNA, 0.5 µL Taq polymerase and nuclease-free water up to 50 µL. To amplify the TEF1α region, two primer pairs and their corresponding PCR conditions were used with both sets amplifying similar regions (Chapter 1 Figure 3).

The primer pair Al33F (5’ GAYTTCATCAAGAACATGAT ’3) and Al33R ((5’ GACGTTGAADCCRACRTTGTC ’3) were used together and Al34F (5’ TTC ATCAAGAACATGAT ’3) and Al34R (5’ GCTATCATCACAATGGACGTTCTTGGAG ’3) were used together. Both primer pairs used the following amplification protocol: 5 minutes of initial denaturing at 94°C, followed by 40 cycles of 50 seconds at 94°C, 50 seconds of annealing at 48°C and 50 seconds at 72°C with a final extension at 72°C for 7 minutes [28]. All PCR reactions were performed in a Sensoquest thermocycler.

2.2.5 Sequencing

PCR products were subjected to electrophoresis in a 1.5% agarose gel containing ethidium bromide and visualised using Ultraviolet illumination. Successful amplification was determined by the presence of a single band at the predicted base pair length. Amplified PCR products were sent for commercial Sanger sequencing by Macrogen Inc. Korea in both forward and reverse directions.

2.2.6 Data Analysis

Sequences returned from Macrogen Inc. Korea were analysed and edited for quality assurance using Sequencher™ ver. 5.2.449 (Gene Codes Corporation). Each nucleotide was given an accuracy confidence score and chromatographs were provided for each sequence. Forward and reverse reads were aligned to form consensus sequences, however, in cases where there was a discrepancy between the forward and reverse reads, confidence scores and chromatographs were observed manually corrected to the nucleotide that was most likely correct. After consensus sequences were formed, Fasta files were exported and verified by aligning the sequences alongside other fungal strains of the same species using CLUSTAL W in MEGA. 7. These alignments were manually checked to verify no obvious

misidentifications or contaminants. In cases where no accurate sequences from the same species were available, members of the same genus were used to check for obvious errors. The sequences were also trimmed at conserved sites to generate sequences with equal 5'- and 3'- endings. These quality-controlled sequences were then submitted to the ISHAM Barcoding Database.

2.3 Results

2.3.1 Barcode Generation

338 fungal strains were subcultured from water cultures of the WM Culture Collection and underwent DNA extraction and barcode amplification (Table 1). After PCR amplification and Sanger sequencing, 313 ITS sequences were generated for 313 fungal strains. These strains represent 49 fungal genera and 100 fungal species. In addition, PCR amplification and Sanger sequencing successfully generated 294 *TEF1α* sequences for 294 fungal strains, representing 93 species and 45 genera. 278 strains had both primary and secondary barcodes generated, 35 strains had only the primary barcode, 19 strains had only the secondary barcode and 6 strains had neither barcode generated (Chapter 2 Figure 1). No obvious trends were present in strains that did not have barcodes generated or where only *TEF1α* sequences were generated. Amongst strains where only ITS sequences were generated, *Naganishia albida* (formerly: *Cryptococcus albidus*), *Vishniacozyma carnescens* (formerly: *Cryptococcus carnescens*), and *Cryptococcus neoformans* var. *grubii* had three or more strains present.

Table 1: Fungal species and the number of strains representing the species with the number of strains with ITS or *TEF1α* sequences generated.

Fungal Species	Number of Strains	Number of Strains with ITS reference sequence generated	Number of Strains with *TEF1α* reference sequence generated
Aspergillus flavus	4	3	4
Aspergillus fumigatus	1	1	1
Aspergillus niger	2	2	2
Aspergillus tubingensis	1	1	1
Blastobotrys adeninivorans	1	1	1
Blastobotrys proliferans	1	1	1
Candida aaseri	1	1	1
Candida albicans	7	7	7
Candida blankii	1	1	1
Candida boleticola	1	1	1
Candida chilensis	1	1	1
Candida coipomensis	1	1	1
Candida diddensiae	2	1	0
Candida diversa	2	2	0
Candida dubliniensis	6	6	6
Candida entomophila	1	1	1
Candida fructus	2	2	2
Candida glabrata	2	2	2
Candida glaebosa	1	1	1
Candida haemulonii (formerly *Candida haemulonis*)	2	2	2
Candida incommunis	2	2	2
Candida inconspicua	7	4	7
Candida insectalens	1	1	1
Candida insectorum	1	1	1
Candida intermedia	2	2	2
Candida membranifaciens	1	1	1
Candida mesenterica	1	1	1
Candida metapsilosis	2	2	1
Candida mucifera	1	1	1
Candida multigemmis	1	1	1
Candida norvegica	1	0	1
Candida orthopsilosis	3	3	3
Candida parapsilosis	15	15	15
Candida railenensis	1	1	1
Candida santamariae	2	2	2
Candida savonica	2	2	2
Candida sequanensis	1	1	1
Candida silvatica	1	0	1
Candida sophiae-reginae	2	2	2
Candida tropicalis	6	4	6

Candida zeylanoides	3	3	3
Citeromyes matritensis	1	1	1
Cladosporium sphaerospermum	1	1	1
Clavispora lusitaniae	5	5	5
Cryptococcus gattii	52	49	50
Cryptococcus neoformans var. *grubii*	17	14	12
Cryptococcus neoformans var. *neoformans*	5	4	4
Cutaneotrichosporon curvatum (formerly *Cryptococcus curvatus*)	1	1	1
Cyberlindnera jadinii	1	1	1
Danielozyma litseae	1	1	0
Danielozyma ontarioensis	1	1	1
Diutina catenulata	1	1	1
Eurotium amstelodami	1	1	1
Filobasidium magnum (formerly *Cryptococcus magnus*)	5	5	3
Fusarium proliferatum	1	1	1
Fusarium solani	1	1	1
Geotrichum candidum	5	5	4
Groenewaldozyma salmanticensis (formerly *Candida salmanticensis*)	1	1	1
Hannaella luteola	1	1	1
Hyphopichia burtonii (formerly *Pichia burtonii*)	1	0	0
Hyphopichia homilentoma (formerly *Candida homilentoma*)	1	1	1
Hyphopichia rhagii (formerly *Candida rhagii*)	1	1	1
Inopinatum lactosum (formerly *Sporobolomyces lactosus*)	1	1	1
Kluyveromyces lactis	1	1	1
Kluyveromyces marxianus	9	9	9
Kodamaea ohmeri	6	5	6
Kuraishia capsulata	1	1	1
Kurtzmaniella fragi (formerly *Candida fragi*)	1	1	1
Magnusiomyces capitatus	5	5	5
Metschnikowia hawaiiana (formerly *Candida hawaiiana*)	1	1	1

Metschnikowia reukaufii	1	1	0
Meyerozyma guilliermondii	6	6	6
Millerozyma farinosa	2	2	2
Naganishia albida (formerly *Cryptococcus albidus*)	6	6	3
Nakazawaea ernobii	2	2	2
Nakazawaea holstii	1	1	1
Papiliotrema aurea (formerly *Cryptococcus aureus*)	1	1	1
Papiliotrema laurentii (formerly *Cryptococcus laurentii*)	2	2	0
Penicillium brevicompactum	1	1	1
Penicillium chrysogenum	1	1	1
Pichia kudriavzevii	7	7	6
Pichia membranifaciens	3	2	3
Rhodotorula mucilaginosa	3	3	3
Saturnispora diversa (formerly *Candida diversa*)	2	2	0
Scedosporium apiospermum	1	0	0
Scedosporium aurantiacum	7	5	5
Scedosporium boydii	2	2	2
Scheffersomyces ergatensis	1	1	0
Starmerella apicola (formerly *Candida apicola*)	1	1	1
Suhomyces canberraensis (formerly *Candida canberraensis*)	1	1	1
Suhomyces prunicola (formerly *Candida prunicola*)	1	1	1
Suhomyces tanzawaensis (formerly *Candida tanzawaensis*)	1	1	1
Torulaspora delbrueckii	2	2	2
Trichomonascus ciferrii	2	2	2
Trichophyton rubrum	3	1	0
Trichosporon asahii	3	3	2
Vanrija humicola (formerly *Asterotremella humicola*)	1	1	0
Vishniacozyma carnescens (formerly *Cryptococcus carnescens*)	8	8	4
Wickerhamomyces anomalus	14	14	12
Yamadazyma mexicana	3	3	3
Yamadazyma scolyti	3	3	3
Yarrowia galli (formerly *Candida galli*)	1	1	1
Yarrowia lipolytica	18	18	17

Zygoascus hellenicus	1	1	1

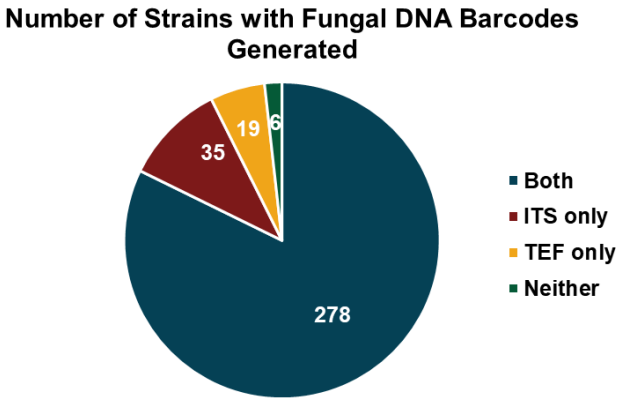


Figure 1: Number of strains with fungal DNA barcodes generated. A total of 313 and 297 ITS and *TEF1α* sequences were generated, respectively from 338 fungal strains. 6 strains had neither the ITS nor the *TEF1α* region amplified. 278 fungal strains had both reference barcodes generated, 35 strains had only the primary barcode generated, and 19 strains had only the secondary barcode generated.

2.3.2 Updated ISHAM fungal DNA barcode Database

After the addition of sequences generated in this study, the ISHAM fungal DNA Database currently holds 4774 primary fungal DNA barcodes and 2432 secondary fungal DNA barcodes. The ITS reference sequences represent 645 human/animal pathogenic fungal species. 275 species are represented by one strain, 124 species are represented by two strains, and 246 species are represented by a minimum of three to a maximum of 155 strains (Chapter 2 Figure 2A). The *TEF1α* reference sequences represent 346 human/animal pathogenic fungal species. 176 species are represented by one strain, 62 species are represented by two strains, and 108 species are represented by a minimum of three to a maximum of 114 strains (Chapter 2 Figure 2B). Additionally, the distribution of strains per species indicated the

species with few strains were more frequent than those with many strains for both the primary and secondary DNA barcodes (Chapter 2 Figure 3).

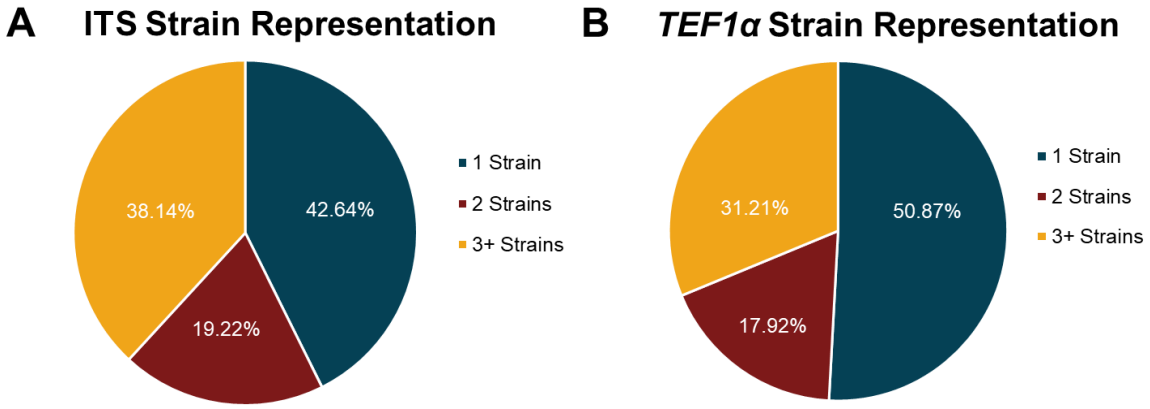


Figure 2: Current percentage distribution of strain representation of species in the ISHAM Barcoding Database divided by 1 Strain, 2 Strains, and 3+ Strains (A) The percentage of species represented by ITS sequences from one strain is 42.64%, from two strains is 19.22% and from a minimum of three strains is 38.14%. **(B)** The percentage of species represented by *TEF1α* sequences from one strain is 50.87%, from two strains is 17.92% and a minimum of three strains is 31.21%.

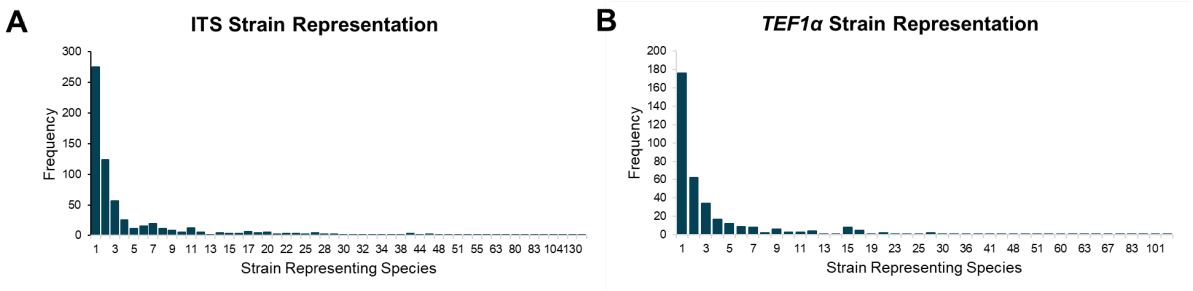


Figure 3: Current frequency distribution of strains per species in the ISHAM Barcoding Database (A) Distribution for the primary fungal DNA barcode, ITS ranges from one strain to 155 strains **(B)** Distribution of the secondary fungal DNA barcode, *TEF1α* ranges from one strain to 114 strains.

2.4 Discussion

DNA barcoding is an important tool for the reliable, rapid, and accurate identification of fungal species in diagnostic settings. The ITS region was commonly used in taxonomic

studies before its establishment as the primary fungal DNA barcode. Its selection was based on ease of amplification and its ability to correctly identify most fungal species. However, it is unable to accurately identify all medically relevant fungal species. As such, *TEF1α* was introduced as the secondary fungal DNA barcode, primarily due to its amplification success rate in a wide array of fungal species and its discriminatory power. The combination of the non-protein coding ITS region and the protein coding *TEF1α* region enables DNA barcoding to discriminate between fast and slow-evolving fungi to provide accurate species identification. Routine use of DNA barcoding in fungal diagnostics is reliant on the availability of robust reference databases containing accurate sequences with high taxonomic coverage. Additionally, high species coverage, meaning more strains representing a species, increases the accuracy of identification as this would more correctly reflect the diversity within a species.

338 fungal strains were prepared for primary and secondary barcode generation in this study. Six strains did not have the primary or secondary barcode generated as amplification was not successful. Culturing, extraction, and PCR amplification were repeated twice for each strain of these six strains and PCR amplification was still unsuccessful. This indicated that there was a problem with the source culture which was then indicated within the culture database. The generation of 313 reference primary fungal barcodes confirmed the high taxonomic coverage achieved by amplification of the ITS region with the universal primer pair ITS1 and ITS4, and the fungal-specific primer pair SR6R and LR1 [29]. As no trends were observed amongst strains where the ITS region was unable to be amplified but the *TEF1α* region was amplified, and so potential impurities left behind from DNA extraction interfered during ITS amplification but not TEF1α amplification. Alternatively, there may have been issues with the specificity of the primers used and so different primer pairs, such as those established in Vigalys and Hester, 1990 or White *et al.* 1990, may be used to amplify the ITS region in

future works [26, 27]. The generation of 294 reference secondary fungal barcodes similarly confirmed the high taxonomic coverage achieved by the *TEF1α* amplifying primers AL33F and AL33R, and Al34F and Al34R. Amongst the 35 strains where the primary barcode was generated but not the secondary barcode, *Naganishia albida* (formerly: *Cryptococcus albidus*), *Vishniacozyma carnescens* (formerly: *Cryptococcus carnescens*), and *Cryptococcus neoformans* var. *grubii,* had three or more strains present. This indicates that there is most likely an issue with the specificity of the primers, as it is unlikely that all amplification failures can be attributed to errors in DNA extraction. *TEF1α* is a protein coding gene and although these regions have high variation between species, PCR amplification has traditionally been challenging, as there is also variation within the primer binding sites making the design of universal primers complicated, due to the lack of conserved regions for primer binding [28]. The selection of *TEF1α* as the secondary fungal DNA barcode had largely been based on amplification success and high taxonomic coverage, which is reflected in the 294 fungal strains with successful *TEF1α* amplification. However, it is evident that the designed primers are not suitable for all fungal species. As such, to increase the taxonomic coverage of the secondary barcode further specific primers should be designed to target the fungal species not amplified by the current primers. The implementation of such primers would benefit the use of DNA barcoding in clinical diagnosis and provide more accurate fungal identification.

The establishment of the secondary fungal DNA barcode was an initiative to improve the accuracy of DNA barcoding by allowing accurate identification of fungal species previously not possible with the primary fungal DNA barcode. Alongside this study, the number of *TEF1α* sequences contained within the ISHAM Database has grown to 2432, far less than the 4774 ITS sequences. The herein generated *TEF1α* sequences now present in the database have established the basis for the secondary fungal DNA database, subsequent barcode

generation is required to bridge the gap between the number of primary and secondary reference barcode sequences in the ISHAM-ITS Database.

The ISHAM-ITS Database has improved upon its first announcement from containing ITS sequences representing 421 species to 645 species and *TEF1α* sequences representing zero species to 346 species. The increase in species coverage is partially due to nomenclature changes and progression in fungal taxonomy studies, however, the continued generation and submission of reference barcodes has mainly driven this growth. The expansion in species coverage of the ISHAM Barcoding database is essential in the functionality of the database to identify all potential pathogenic fungal species. As such, continued work to generate reference barcode sequences for all pathogenic species is required for both primary and secondary barcodes, and further work to include new pathogenic fungal species is necessary.

Aside from the number of species present in the database, the number of strains and thereby reference sequences representing each species is equally essential to accurately reflect the diversity within species, thereby providing more accurate identifications. Currently, 57.28% and 54.89% of fungal species within the ISHAM database have multiple primary and secondary barcodes attributed to a single species, respectively. As such, 42.72% and 45.11% of fungal species are only represented by one strain for the primary and secondary barcodes, respectively (Figure 3). This discrepancy is evident in the hyperbolic distribution of strains per species (Figure 4). These values are consistent with the initial establishment of the ISHAM-ITS Database, of which 41.81% of fungal species were represented by one strain. A more concentrated effort is required to reduce the proportion of fungal species with a reference sequence of only one strain. However, these species are largely rare proponents of disease as such barcode generation from clinical samples containing these species is challenging. To overcome this, strains from environmental sources belonging to the same fungal species should be considered for barcode generation. Additionally, sequences from

other databases that undertake quality-control measures may also provide further sequences for underrepresented fungal species.

2.5 Conclusion

The generation of accurate reference barcodes and the organisation of such sequences in trusted databases with rigorous quality control measures are essential for the use of DNA barcoding in the diagnosis of mycoses. Further work is required to increase the taxonomic coverage for both fungal DNA barcodes to encompass all pathogenic fungal species and to increase species representation. Besides these issues, the ISHAM-ITS DNA barcode database is the most quality controlled fungal DNA barcode database and has continued its growth since its establishment in 2015, including the additional 610 reference barcode sequences generated in this study.

2.6 References

1. Irinyi, L., *et al.*, *International Society of Human and Animal Mycology (ISHAM)-ITS reference DNA barcoding database—the quality controlled standard tool for routine identification of human and animal pathogenic fungi.* Med. Mycol., 2015. **53**(4): p. 313-337.
2. Karsch-Mizrachi, I., *et al.*, *The international nucleotide sequence database collaboration.* Nucleic Acids Res., 2018. **46**(D1): p. D48-D51.
3. Sayers, E.W., *et al.*, *GenBank.* Nucleic Acids Res., 2020. **48**(D1): p. D84-D86.
4. Sayers, E.W. and I. Karsch-Mizrachi, *Using GenBank.* Methods Mol. Biol., 2016. **1374**: p. 1-22.

5. Ryberg, M., *et al.*, *An outlook on the fungal internal transcribed spacer sequences in GenBank and the introduction of a web-based tool for the exploration of fungal diversity.* New Phytol., 2009. **181**(2): p. 471-477.

6. Steinegger, M. and S.L. Salzberg, *Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank.* Genome Biol., 2020. **21**(1): p. 115.

7. Locatelli, N.S., *et al.*, *GenBank's reliability is uncertain for biodiversity researchers seeking species-level assignment for eDNA.* Proc. Natl. Acad. Sci. U. S. A., 2020. **117**(51): p. 32211-32212.

8. Meiklejohn, K.A., N. Damaso, and J.M. Robertson, *Assessment of BOLD and GenBank - Their accuracy and reliability for the identification of biological materials.* PLoS One, 2019. **14**(6): p. e0217084.

9. Pentinsaari, M., *et al.*, *BOLD and GenBank revisited - Do identification errors arise in the lab or in the sequence libraries?* PLoS One, 2020. **15**(4): p. e0231814.

10. Nilsson, R.H., *et al.*, *Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective.* PLoS One, 2006. **1**: p. e59.

11. Ratnasingham, S. and P.D. Hebert, *bold: The Barcode of Life Data System (http://www.barcodinglife.org).* Mol. Ecol. Notes, 2007. **7**(3): p. 355-364.

12. Ratnasingham, S. and P.D. Hebert, *A DNA-based registry for all animal species: the barcode index number (BIN) system.* PLoS One, 2013. **8**(7): p. e66213.

13. deWaard, J.R., *et al.*, *A reference library for Canadian invertebrates with 1.5 million barcodes, voucher specimens, and DNA samples.* Sci. Data, 2019. **6**(1): p. 308.

14. Pruitt, K.D., T. Tatusova, and D.R. Maglott, *NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.* Nucleic Acids Res., 2007. **35**(Database issue): p. D61-5.

15. O'Leary, N.A., *et al.*, *Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.* Nucleic Acids Res., 2016. **44**(D1): p. D733-45.

16. Sayers, E.W., *et al.*, *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res., 2020. **48**(D1): p. D9-D16.

17. Schoch, C.L., *et al.*, *Finding needles in haystacks: linking scientific names, reference specimens and molecular data for Fungi.* Database (Oxford), 2014. **2014**: p. 1-21.

18. Koljalg, U., *et al.*, *UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi.* New Phytol., 2005. **166**(3): p. 1063-8.

19. Nilsson, R.H., *et al.*, *Molecular identification of fungi: Rationale, philosophical concerns, and the UNITE database.* The Open Applied Informatics Journal, 2011. **5**(1): p. 81-86.

20. Koljalg, U., *et al.*, *Towards a unified paradigm for sequence-based identification of fungi.* Mol. Ecol., 2013. **22**(21): p. 5271-7.

21. Abarenkov, K., *et al.*, *The UNITE database for molecular identification of fungi - recent updates and future perspectives.* New Phytol., 2010. **186**(2): p. 281-5.

22. Tedersoo, L., *et al.*, *Tidying up international nucleotide sequence databases: ecological, geographical and sequence quality annotation of its sequences of mycorrhizal fungi.* PLoS One, 2011. **6**(9): p. e24940.

23. Nilsson, R.H., *et al.*, *The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications.* Nucleic Acids Res., 2019. **47**(D1): p. D259-D264.

24. Irinyi, L., *et al.*, *DNA barcoding of fungi causing infections in humans and animals.* Fungal Biol., 2016. **120**(2): p. 125-36.

25. Ferrer, C., *et al.*, *Detection and identification of fungal pathogens by PCR and by ITS2 and 5.8 S ribosomal DNA typing in ocular infections.* J. Clin. Microbiol., 2001. **39**(8): p. 2873-2879.

26. Vilgalys, R. and M. Hester, *Rapid genetic identification and mapping of enzymatically amplified ribosomal DNA from several Cryptococcus species.* J. Bacteriol., 1990. **172**(8): p. 4238-46.

27. White, T., J., *et al.*, *Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics*, in *PCR protocols: A guide to methods and applications*. 1990, Academic Press: New York. p. 315-322.

28. Stielow, J.B., *et al.*, *One fungus, which genes? Development and assessment of universal primers for potential secondary fungal DNA barcodes.* Persoonia, 2015. **35**: p. 242.

29. Schoch, C.L., *et al.*, *Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for fungi.* Proc. Natl. Acad. Sci. U. S. A., 2012. **109**(16): p. 6241-6246.

Chapter 3
Database Establishment for the Secondary Fungal DNA Barcode

3.1 Database establishment for the secondary fungal DNA barcode *translational elongation factor 1α* (*TEF1α*)

This manuscript was published in Genome in November 2018 (volume 62, issue 3, pages 160-169). Hoang M.T.V. was involved in performing the research and analysis of data and was involved in the revision of the manuscript. Meyer W. is the primary author and contributed to all aspects of this study. Irinyi L. was involved in performing the research, analyzing the data, and preparing the manuscript. All authors were involved in the revision of the manuscript and support its inclusion in this thesis.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Minh Thuy Vi Hoang
30th June 2023

As a supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Professor Wieland Meyer
30th June 2023

NRC Research Press

ARTICLE

Database establishment for the secondary fungal DNA barcode *translational elongation factor 1α (TEF1α)*[1]

Wieland Meyer, Laszlo Irinyi, Minh Thuy Vi Hoang, Vincent Robert, Dea Garcia-Hermoso, Marie Desnos-Ollivier, Chompoonek Yurayart, Chi-Ching Tsang, Chun-Yi Lee, Patrick C.Y. Woo, Ivan Mikhailovich Pchelin, Silke Uhrlaß, Pietro Nenoff, Ariya Chindamporn, Sharon Chen, Paul D.N. Hebert, Tania C. Sorrell, and the ISHAM barcoding of pathogenic fungi working group[2]

Genome Downloaded from cdnsciencepub.com by University of Sydney on 05/03/21
For personal use only.

Abstract: With new or emerging fungal infections, human and animal fungal pathogens are a growing threat worldwide. Current diagnostic tools are slow, non-specific at the species and subspecies levels, and require specific morphological expertise to accurately identify pathogens from pure cultures. DNA barcodes are easily amplified, universal, short species-specific DNA sequences, which enable rapid identification by comparison with a well-curated reference sequence collection. The primary fungal DNA barcode, ITS region, was introduced in 2012 and is now routinely used in diagnostic laboratories. However, the ITS region only accurately identifies around 75% of all medically relevant fungal species, which has prompted the development of a secondary barcode to increase the resolution power and suitability of DNA barcoding for fungal disease diagnostics. The translational elongation factor 1α (*TEF1α*) was selected in 2015 as a secondary fungal DNA barcode, but it has not been implemented into practice, due to the absence of a reference database. Here, we have established a quality-controlled reference database for the secondary barcode that together with the ISHAM-ITS database, forms the ISHAM barcode database, available online at http://its.mycologylab.org/. We encourage the mycology community for active contributions.

Key words: fungal DNA barcoding, secondary fungal DNA barcode database, translational elongation factor 1α.

Résumé : En raison d'infections fongiques nouvelles ou émergentes, les champignons pathogènes humains et animaux constituent une menace croissante mondialement. Les outils diagnostiques actuels sont lents, non-discriminants au niveau spécifique ou sous-spécifique, et ils nécessitent une expertise morphologique pour identifier avec justesse les agents pathogènes en cultures pures. Les codes à barres de l'ADN sont des séquences

Received 9 May 2018. Accepted 5 November 2018.

Corresponding Editor: Jianping Xu.

W. Meyer, L. Irinyi, M.T.V. Hoang, and T.C. Sorrell. Molecular Mycology Research Laboratory, Centre for Infectious Diseases and Microbiology, Faculty of Medicine and Health, Sydney School of Medicine, Westmead Clinical School, Marie Bashir Institute for Infectious Diseases and Biosecurity, The University of Sydney, Westmead Hospital (Research and Education Network), Westmead Institute for Medical Research, Westmead, NSW, Australia.
V. Robert. Westerdijk Fungal Biodiversity Institute, Utrecht, the Netherlands.
D. Garcia-Hermoso and M. Desnos-Ollivier. Institut Pasteur, National Reference Center for Invasive Mycoses and Antifungals (NRCMA), Molecular Mycology Unit, CNRS UMR2000, Paris, France.
C. Yurayart. Mycology Unit, Department of Microbiology, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand; Department of Veterinary Microbiology and Immunology, Faculty of Veterinary Medicine, Kasetsart University, Bangkok, Thailand.
C.-C. Tsang, C.-Y. Lee, and P.C.Y. Woo. Department of Microbiology, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong.
I.M. Pchelin. Laboratory of Molecular Genetic Microbiology, Kashkin Research Institute of Medical Mycology, I.I. Mechnikov North-Western State Medical University, St Petersburg, Russia.
S. Uhrlaß and P. Nenoff. Laboratory of Medical Microbiology, Partnership Dr. C. Krueger & Prof. Dr. P. Nenoff, Roetha OT Moelbis, Germany.
A. Chindamporn. Mycology Unit, Department of Microbiology, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand.
S. Chen. Molecular Mycology Research Laboratory, Centre for Infectious Diseases and Microbiology, Faculty of Medicine and Health, Sydney School of Medicine, Westmead Clinical School, Marie Bashir Institute for Infectious Diseases and Biosecurity, The University of Sydney, Westmead Hospital (Research and Education Network), Westmead Institute for Medical Research, Westmead, NSW, Australia; Centre for Infectious Diseases and Microbiology Laboratory Services, ICPMR, Westmead Hospital, Westmead, NSW, Australia.
P.D.N. Hebert. Department of Integrative Biology and Director of the Biodiversity Institute of Ontario at the University of Guelph, Guelph, ON, Canada.

Corresponding author: Wieland Meyer (email: wieland.meyer@sydney.edu.au).

[1]This paper is part of a special issue entitled "Trends in DNA Barcoding and Metabarcoding".

[2]See Acknowledgments for complete list of working group members.

Copyright remains with the author(s) or their institution(s). Permission for reuse (free in most cases) can be obtained from RightsLink.

Genome **62**: 160–169 (2019) dx.doi.org/10.1139/gen-2018-0083

Published at www.nrcresearchpress.com/gen on 22 November 2018.

d'ADN courtes, spécifiques, faciles à amplifier et universelles, ce qui rend possible une identification rapide par simple comparaison avec une collection de séquences de référence bien constituée. Le principal code à barres de l'ADN chez les champignons est la région ITS, laquelle a été employée initialement en 2012 et est maintenant employée de manière routinière dans les laboratoires diagnostiques. Cependant, la région ITS ne permet d'identifier avec justesse qu'environ 75 % de tous les spécimens fongiques d'intérêt médical, ce qui a créé la nécessité de développer un code à barres secondaire de manière à augmenter la résolution et la pertinence du codage à barres de l'ADN pour le diagnostic des maladies fongiques. Le facteur d'élongation de la traduction 1α (*TEF1α*) a été choisi en 2015 comme code à barres secondaire chez les champignons, mais n'a pas été tellement employé en raison de l'absence d'une base de données de séquences de référence. Dans ce travail, les auteurs ont établi une telle base de référence dont la qualité a été vérifiée pour le code à barres secondaire qui, conjointement à la base de données ISHAM-ITS, forme la base de données « ISHAM barcode database », disponible en ligne à http://its.mycologylab.org/. Les auteurs encouragent la communauté des mycologues à y contribuer activement. L'utilisation d'un système de codage à barres double facilitera l'identification juste de tous les champignons pathogènes d'intérêt clinique.

Mots-clés : codage à barres de l'ADN fongique, base de données d'un code à barres fongique secondaire, facteur d'élongation de la traduction 1α.

Genome Downloaded from cdnsciencepub.com by University of Sydney on 05/03/21
For personal use only.

Introduction

The identification and delimitation of fungal species are often challenging using traditional methods, either because of the lack of distinct morphological and biochemical characters or the absence of sexual reproduction. Complex life cycles, such as yeast–mycelial transitions or the existence of synanamorphs of certain species further hamper correct morphological identification (Seifert and Samuels 2000). Conventional fungal identification methods using culture and microscopic analyses are often insensitive, slow (7–14 days), and heavily dependent on the level of mycological expertise (Alexander and Pfaller 2006; Balajee et al. 2007). Further morphological and biochemical traits are subjective, error prone, and frequently result in misidentifications (Sangoi et al. 2009). In addition, the fact that numerous fungal species cannot grow under laboratory conditions precludes their identification (Begerow et al. 2010; Nilsson et al. 2009). Phenotypic traits also fail to differentiate closely related species or species complexes with near-identical morphological characters but distinguishable genetic traits, such as, e.g., the members of the genera *Fusarium*, *Aspergillus*, and *Scedosporium* (Balajee et al. 2005; Gilgado et al. 2005; O'Donnell 2000).

Rapid and accurate identification is essential in clinical diagnostics, outbreak investigation, or other epidemiological studies, as well as in agricultural, quarantine, and industrial settings. Culture and "expert-free" methods capable of identifying fungi directly from biological specimens are needed. The early diagnosis of pathogens is critical in fungal disease management, public health, animal welfare, and plant protection, for quarantine purposes and in other industrial applications. To counter these difficulties in identification, many molecular techniques have been developed over the last 20 years. These include PCR-RFLP analysis (Dendis et al. 2003; Diguta et al. 2011), RAPD (Baires-Varguez et al. 2007; Brandt et al. 1998), hybridisation with genus/species specific DNA or RNA probes (Lindsley et al. 2001; Sandhu et al. 1995), PCR fingerprinting (Hierro et al. 2004; Lieckfeldt et al. 1993), species-specific PCR assays (Kulik et al. 2004; Martin et al. 2000), real time PCR (Bergman et al. 2007; Klingspor and Jalal 2006), and, increasingly, DNA sequencing (Pryce et al. 2003; Romanelli et al. 2010). Sequence-based identification methods have proven to be more accurate than conventional methods in diagnostic clinical mycology (Balajee et al. 2007; Ciardo et al. 2006). Developments in molecular-based technologies are expected to improve further and simplify species or even strain identification, making it faster and more accurate in the future.

Among the applied molecular techniques, DNA barcoding is one of the most promising and efficient methods, as it enables rapid identification of species and recognition of cryptic species across all kingdoms of eukaryotic life. The concept was first proposed by Hebert and colleagues in 2003 (Hebert et al. 2003*a*). Barcodes are standardized, easily amplified, universal short DNA sequences (500–800 bp), which exhibit a high divergence at species level and allow rapid identification by comparison with a reference sequence collection of accurately identified species (Hajibabaei et al. 2007). Characteristics of barcodes necessary for their routine use include the following: ease of attainment from any specimen, irrespective of morphology or stage of the fungal life cycle; high taxonomic coverage; and a high resolution power (Hebert et al. 2003*a*). Ideally, barcodes must be unique to a single species, and to be constant within each species to ensure consistency of identification (Hebert et al. 2003*a*, 2003*b*; Letourneau et al. 2010). Additionally, interspecies variation should exceed the intraspecies variation, generating a "break" in the distribution of distances that is referred to as the barcoding gap (Meyer and Paulay 2005). This gap is generally evaluated and represented by the Kimura 2-parameter distance model (Kimura 1980).

Numerous genetic loci have been evaluated as a barcode for fungi, with varying success rates. Protein-coding genes show high taxonomic resolution and identification success in Ascomycota. The most commonly used

Genome Downloaded from cdnsciencepub.com by University of Sydney on 05/03/21
For personal use only.

genes are *RPB1* (Crespo et al. 2007; Hofstetter et al. 2007; McLaughlin et al. 2009; O'Donnell et al. 2013; Tanabe et al. 2005), *β-tubulin* (Frisvad and Samson 2004), and partial *translational elongation factor 1α* (*TEF1α*) (James et al. 2006; O'Donnell et al. 2010; Schoch et al. 2009). The D1/D2 region of the large subunit (LSU) of the rDNA gene cluster has been used successfully for species identification and strain recognition in yeasts (Fell et al. 2000; Kurtzman and Robnett 1998; Scorzetti et al. 2002). However, the applicability of these genetic loci as a potential barcode is hampered by the lack of standardization of the amplification process, differences in primer selection, their species/genera-dependent resolution power, and the lack of quality-controlled reference databases.

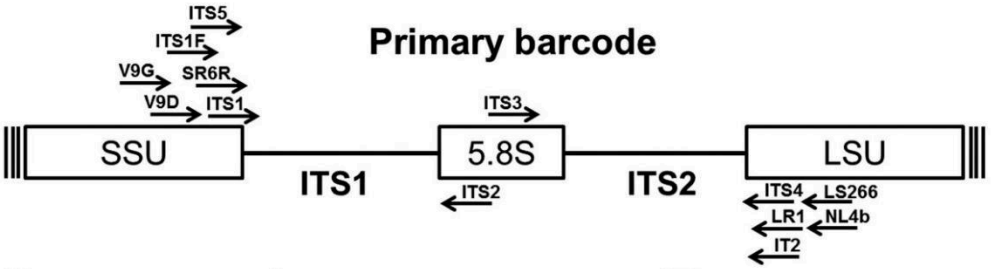
The primary fungal barcode

The internal transcribed spacer (ITS) region is comprised of three units, two non-coding, variable, ITS1 and ITS2 regions, separated by the highly conserved 5.8S gene, which is located between the 18S (small subunit (SSU)) and 28S (LSU) genes in the nrDNA repeat unit (White et al. 1990). The ITS region has long been used for species identification and phylogenetic studies in fungi (Bruns et al. 1991; Gardes and Bruns 1993; Hillis and Dixon 1991; Leaw et al. 2006; Meyer and Gams 2003). Its widespread popularity relies on several facts. There are many universal primers available to bind to DNA of most of the fungal taxa, with the most commonly used being the ITS1, ITS1F, ITS2, ITS3, ITS4, and ITS5 (Gardes and Bruns 1993; White et al. 1990). To avoid cross reactivity with plant or animal DNA, fungus-specific primers were also designed, such as SR6R and LR1 (Vilgalys and Hester 1990), V9D, V9G, and LS266 (Gerrits van den Ende and de Hoog 1999), IT2 (Beguin et al. 2012), ITS1F (Gardes and Bruns 1993), and NL4b (O'Donnell 1993) (Fig. 1). Multiple copies of the ITS are present in the genome which increases significantly the amplification efficiency, even from samples where the initial amount of DNA is low, such as environmental or clinical specimens (Vilgalys and Gonzalez 1990). The relatively short length of the ITS region allows easy amplification and subsequent Sanger sequencing (Seifert 2009). Finally, the ITS region has proven to have a good resolution power leading to species discrimination in most fungal taxa. A significant drawback is its limited resolution in closely related species, or species complexes containing cryptic species (Nilsson et al. 2008). Recognising the many advantages and relatively few limitations of the ITS region for use in identification across the fungal kingdom, it was proposed as a standard fungal DNA barcode in 2012 by Schoch and colleagues (Schoch et al. 2012) — although there were strong criticisms of this proposal by some mycologists (Kiss 2012). The main shortcoming of barcoding for microorganisms, not limited to fungi, has been the lack of adequate and validated reference libraries of DNA barcodes (Seifert 2009). In response to this, the ISHAM-ITS reference database for human and animal pathogenic fungi was established in 2015 (Irinyi et al. 2015). Currently (as of 21 September 2018) this database is in global use and contains 4200 complete ITS sequences, representing 645 fungal species. The ITS locus proved to be sufficient for correct identification of most medically relevant fungal species (Irinyi et al. 2015). However, an exact cut-off point for identification could not be defined, in agreement with other studies (Nilsson et al. 2008), due to a high number of polymorphisms (up to 2.5% of the ITS region) in certain species. For these taxa, it is essential to consider alternative genetic regions to obtain superior resolution power and separation. Correct identification of these taxa cannot be guaranteed by using only ITS sequences, as has been documented in the global project "Assembling the Fungal Tree of Life" (https://aftol.umn.edu) (James et al. 2006). This is particularly relevant to species complexes and sibling species delineation. Preferably, an alternative barcode should contain a single, probably protein-coding gene sequence with high interspecies and low intraspecies divergence and high taxonomic coverage, hence meeting the requirements of a standard barcode, namely, amplified by universal primers with high PCR efficiencies.

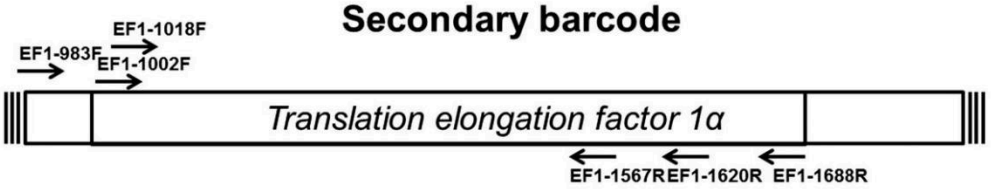
The secondary fungal barcode

Despite the advantages of the ITS locus, such as robust PCR amplification and a high taxonomic coverage, it can only accurately identify around 75% of all medically relevant fungal species using a cut-off of 98.5% sequence identity (Irinyi et al. 2015). It has also been shown that the resolution power of ITS at a higher taxonomic level is inferior to that of many protein coding genes (Nilsson et al. 2006; Seifert 2009). To enhance the selectivity of fungal DNA barcoding and enable the correct identification of all fungal species, the most pragmatic solution was to establish a secondary fungal barcode, recognising that it would be difficult to identify only one DNA region that fulfilled all requirements of a barcode (Stielow et al. 2015). Over the years, many loci have been tested as alternatives to the ITS region but no consensus had been reached among mycologists mainly due to the lack of standardization and the non-availability of universal primers (Frisvad and Samson 2004; McLaughlin et al. 2009; O'Donnell et al. 2010). Whole genome sequencing (WGS) has recently made it possible to carry out full genome comparisons of evolutionarily distinct fungi to identify potential secondary barcode markers and design universal primers for robust PCR amplification (Robert et al. 2011). The proposal was to identify genetic markers with high interspecies and low intraspecies sequence divergence, which accurately reflect higher-level taxonomic affiliations. In a recent study, 14 (partially) universal primer pairs targeting eight genetic markers were tested across more than 1500 species (1931 strains or specimens) to select the most optimal secondary barcode

Fig. 1. Schematic structure of the primary (ITS) and secondary (*TEF1α*) fungal DNA barcode regions indicating universal primers for their amplification.



Primer name	Sequences	References
ITS1:	(5' TCCGTAGGTGAACCTGCGG 3')	(White et al., 1990)
ITS2:	(5' GCTGCGTTCTTCATCGATGC 3')	(White et al., 1990)
ITS3:	(5' GCATCGATGAAGAACGCAGC 3')	(White et al., 1990)
ITS4:	(5' TCCTCCGCTTATTGATATGC 3')	(White et al., 1990)
ITS5:	(5' GGAAGTAAAAGTCGTAACAAGG 3')	(White et al., 1990)
ITS1F:	(5' CTTGGTCATTTAGAGGAAGTAA 3')	(Gardes and Bruns, 1993)
IT2:	(5' CCTCCGCTTATTGATATGCTTAGG 3')	(Beguin et al., 2012)
SR6R:	(5' AAGTATAAGTCGTAACAAGG 3')	(Vilgalys and Hester, 1990)
LR1:	(5' GGTTGGTTTCTTTCCT 3')	(Vilgalys and Hester, 1990)
V9D:	(5' TTAAGTCCCTGCCCTTTGTA 3')	(Gerrits van den Ende and de Hoog, 1999)
V9G:	(5' TACGTCCCTGCCCTTTGTA 3')	(Gerrits van den Ende and de Hoog, 1999)
LS266:	(5' GCATTCCCAAACAACTCGACTC 3')	(Gerrits van den Ende and de Hoog, 1999)
NL4b:	(5' GGATTCTCACCCTCTATGAC 3')	(O'Donnell, 1993)



Primer name	Sequences	References
EF1-1018F (Al33F):	(5' GAYTTCATCAAGAACATGAT 3')	(Stielow et al., 2015)
EF1-1620R (Al33R):	(5' GACGTTGAADCCRACRTTGTC 3')	(Stielow et al., 2015)
EF1-1002F (Al34F):	(5' TTCATCAAGAACATGAT 3')	(Stielow et al., 2015)
EF1-1688R (Al34R):	(5' GCTATCATCACAATGGACGTTCTTGGAG 3')	(Stielow et al., 2015)
EF1-983F:	(5' GCYCCYGGHCAYCGTGAYTTYAT 3')	(Rehner and Buckley, 2005)
EF1-1567R:	(5' ACHGTRCCRATACCACCRATCTT 3')	(Rehner and Buckley, 2005)

marker (Stielow et al. 2015). As a result of this study the *TEF1α* gene proved to be the most promising candidate, and as such it was proposed as the universal secondary fungal DNA barcode, based on its universal taxon applicability and the availability of universal primers, such as EF1-1018F (Al33F)/EF1-1620R (Al33R) or EF1-1002F (Al34F)/EF1-1688R (Al34R) (Fig. 1).

The aims of this publication are as follows:

- To announce the establishment of a quality controlled reference database for the secondary DNA barcode (*TEF1α*) of human and animal pathogenic fungi as a tool to fill the current gap in the diagnosis of mycoses.
- To make a call to the global medical mycology community for further sequence contributions to enable full coverage of all medically important fungal species.

Materials and methods

Cultures

In this study 908 strains, representing 186 human and animal pathogenic fungal species, were used to generate quality controlled *TEF1α* sequences.

DNA extraction

DNA was isolated and purified from cultures using either previously described methods (Ferrer et al. 2001)

Genome Downloaded from cdnsciencepub.com by University of Sydney on 05/03/21
For personal use only.

Genome Downloaded from cdnsciencepub.com by University of Sydney on 05/03/21
For personal use only.

or the Quick-DNA Fungal/Bacterial Kit (D6007, Zymo Research), according to the manufacturer's instructions. Laboratories contributing to the current study carried out the DNA extractions by using the methods routinely used in their laboratories.

PCR amplification of the secondary barcode

All PCR reactions were performed in a final volume of 50 μL, made of 5 μL dNTP mix, 5 μL 10 × reaction buffer, 4 μL 10 ng/μL forward and reverse primers, 3 μL 50 mM $MgCl_2$, 15 μL 10 ng/μL genomic DNA, 0.5 μL Taq polymerase, and nuclease-free water up to 50 μL. Partial amplification of the secondary fungal DNA barcode, *TEF1α* region, was done using two sets of primers: Al33F (5′-GAYTTCATCAAGAACATGAT-′3) and Al33R (5′-GACGTT GAADCCRACRTTGTC-′3) or Al34F (5′-TTCATCAAGAAC ATGAT-′3) and Al34R (5′-GCTATCATCACAATGGACGTT CTTGGAG-′3) with the following amplification protocol: 5 min initial denaturing at 94 °C, followed by 40 cycles of 50 min at 94 °C, 50 s annealing at 48 °C, 50 s at 72 °C, and 7 min final extension at 72 °C (Stielow et al. 2015). First, the primer set Al33F-Al33R was used to amplify the *TEF1α* region, and if there was no amplification achieved the second primer set Al34F-Al34R was applied. PCR products were subjected to electrophoresis in a 1.5% agarose gel containing EtBr and visualised using UV illumination. Successfully amplified PCR products were sent for commercial sequencing, e.g., to Macrogen Inc., Korea, in both forward and reverse directions.

Data analysis

Bidirectional sequences were assembled and edited using Sequencher® ver. 5.3. (Gene Codes Corporation, Ann Arbor, MI USA). Sequences were manually checked, and ambiguous bases were corrected based on the forward and reverse trace files, considering the PHRED scores received. The sequences for each taxon were aligned with the program CLUSTALW (Thompson et al. 1994), which is part of the software MEGA ver. 5.2.2 (Tamura et al. 2011). Resulting multiple alignments were then checked visually and edited when needed. For further analyses, the sequences were truncated at conserved sites to obtain equal 3′- and 5′-endings. The intraspecies diversity was estimated by calculating the average nucleotide diversity (π) within species, which had sequences from more than three strains, giving the proportion of nucleotide differences in all haplotypes in the studied sample using the software DnaSP ver. 5.10.01 (Librado and Rozas 2009).

Results and discussion

Establishment of the secondary fungal barcode database

With the *TEF1α* locus being selected as secondary barcode, a dedicated quality-controlled database covering all medically relevant fungal species was established to complement the ISHAM-ITS database.

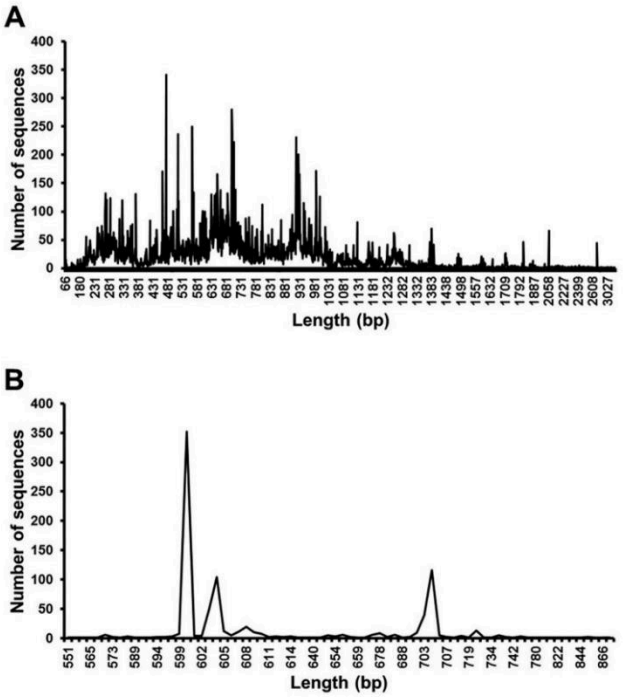
Currently there are 37 773 *TEF1α* sequences available in GenBank (as of 15 March 2018), representing 6283 taxa, of which 670 are only classified to the genus level. However, there are only 6539 sequences from 269 clinically relevant species in 108 fungal genera. The majority of them belong to the genera *Fusarium* (2647), *Cryptococcus* (694), and *Sporothrix* (154). Of these, 114 species are represented by less than two sequences. The length of the *TEF1α* sequences in GenBank ranges from 66 to ~3000 bp (Fig. 2A), indicating that some of the sequences might be too short for correct species identification. In addition, these sequences are not quality controlled and have not been verified at species level, which may lead to an incorrect identification.

The secondary fungal DNA barcode database was established using only quality-controlled *TEF1α* sequences obtained from taxonomically verified fungal cultures. The newly established database contains 908 quality-controlled *TEF1α* sequences, representing 186 pathogenic fungal species, was launched at the recent 7th International DNA Barcoding Conference at the Kruger National Park in South Africa (20–24 November 2017). Sequences from the genera *Candida*, *Cryptococcus*, and *Scedosporium* are most abundant in the new database. In total, 143 species are represented by three strains or less. The length of the partial *TEF1α* sequences in the database ranges from 550–900 bp (Fig. 2B). In a few cases, shorter amplification products of the *TEF1α* were obtained (e.g., for *Trichosporon* spp.).

Overall, the PCR success rate was very good, e.g., in the Molecular Mycology Research Laboratory, Westmead Institute for Medical Research, Westmead, NSW, Australia, 270 *TEF1α* sequences were generated. From those, 220 were produced using the Al33F-Al33R primer set and 50 were amplified with the Al34F-Al34R primer set. There was no trend in amplification success for different fungal species; however, all strains for the species *Aspergillus niger*, *Candida albicans*, *Candida dublinsiensis*, *Kluyveromyces marxianus*, and *Pichia kudriavzevii* required the Al34f-A134R primer set for amplification. However, in some cases no amplification products for some strains of *Cladosporium* spp., *Rhodotorula* spp., or *Trichosporon* spp. were obtained. In those cases, modifications of the amplification conditions, like touchdown PCR, may be appropriate.

Comparison of the nucleotide diversity (π) and number of polymorphic sites (S) of 43 fungal species, with more than three strains per species in the new secondary fungal DNA barcode database (Fig. 3), showed that the *TEF1α* locus is less diverse than the ITS locus for most of the species, with the intraspecies variation being below 1.5%. The pilot data reported herein showed a further reduction in the intraspecies diversity using the *TEF1α* locus compared with the ITS region for 12 fungal species, including *Candida albicans*, *C. glabrata*, *C. metapsilosis*, *C. pararugosa*, *C. tropicalis*, *Clavispora lusitaniae*, *C. gattii*,

Fig. 2. Length distribution of secondary barcode (*TEF1α*) sequences in (A) GenBank and (B) ISHAM barcode database. Only *TEF1α* sequences longer than 550 bp were included.



Genome Downloaded from cdnsciencepub.com by University of Sydney on 05/03/21
For personal use only.

Histoplasma capsulatum, *Kodamaea ohmeri*, *Scedosporium apiospermum*, *S. aurantiacum*, and *Yarrowia lipolytica*. The combination of the increased number of fungal species, which had less than 1.5% intraspecies variation, with the further reduction in intraspecies variation compared to the ITS confirms the secondary DNA barcode locus (*TEF1α*) as a more discriminatory marker. Further studies, including more strains and covering all human and animal pathogenic fungi, are needed to enable specific indications for the use of the dual fungal DNA barcode.

As a result of this study, the ITS and the *TEF1α* databases where combined together to form the new ISHAM barcode database, which currently contains 4200 ITS and 908 *TEF1α* sequences. The ISHAM barcode database is available online at the following websites: http://its.mycologylab.org/ or the ISHAM website (http://isham.org). This new database allows single locus or polyphasic identification of human and animal pathogenic fungi based on sequence alignments of either the ITS or the *TEF1α* locus sequences, or both, against the reference sequences maintained in the ISHAM barcode database.

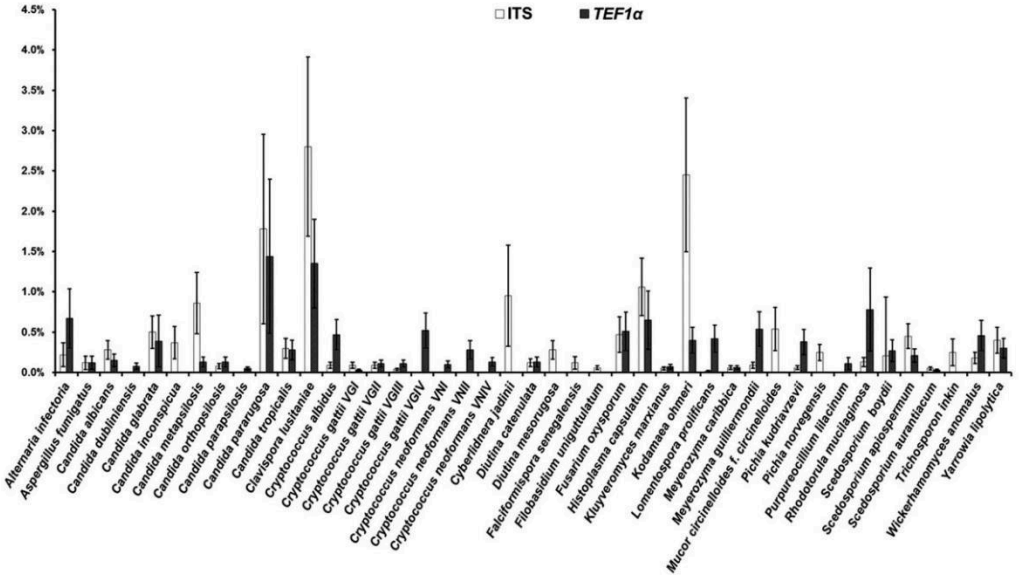
Call for data submission

The database is open for further sequence submissions with the goal of eventually containing sequences from multiple strains of all medically relevant species. Data obtained from taxonomically validated strains can be submitted directly into the database or via contacting the curators of the database: DNA-barcode.database@sydney.edu.au. After validation of the taxonomic metadata and verification of the submitted sequences by the database curator(s), they will be added to the database.

Value of the dual barcoding system for fungal identification

In summary, our data confirm that most medically relevant fungal species can be identified from their ITS sequence, confirming its status as a primary universal fungal DNA barcode. However, some species may not be distinguishable using the ITS region. This can be the case for the following two reasons: either the taxa are not adequately represented in the database, or the ITS region is unable to distinguish between biologically consistent groups. Under these circumstances, the application of a dual DNA barcoding system that incorporates a secondary barcode, *TEF1α*, supported by a quality-controlled reference database, increases the ability to accurate identification of all clinically important fungal pathogens. The herein established secondary fungal DNA barcode

Fig. 3. Intraspecies variation for species, which are represented by more than three strains, in the ITS regions (white bars) — primary DNA barcode — compared with the translation elongation factor 1α (black bars) — secondary DNA barcode.



database offers the basis for a comprehensive database of the fungal kingdom.

Acknowledgements

The authors thank Jeffrey M. Loch (U.S. Geological Survey, National Wildlife Health Center, Madison, USA) for submitting sequences and all reference centres and clinical microbiology laboratories that kindly contributed strains to this study. The study was supported by NH&MRC grants APP1031952 and APP1121936 to W.M., V.R., S.C., and T.C.S., and a REN grant from Western Sydney Local Health Research & Education Network to W.M., T.C.S., S.C., L.I. T.C.S. is a Sydney Medical Foundation Fellow.

Members of the ISHAM barcoding of pathogenic fungi working group include the following: Sharon Chen, Catriona Halliday, Laszlo Irinyi, Wieland Meyer, Tania C. Sorrell (University of Sydney/Westmead Hospital, Sydney, Australia); Ian Arthur (PathWest Laboratory Medicine WA, Nedlands, Australia); Maria Luiza Moretti (University of Campinas, Campinas, Brazil); Célia Maria de Almeida Soares (Universidade Federal de Goiás, Goiânia, Brazil); Mauro de Medeiros Muniz, Rosely Maria Zancopé-Oliveira (Instituto de Pesquisa Clínica Evandro Chagas (IPEC), Fundação Oswaldo Cruz (FIOCRUZ), Rio de Janeiro, Brazil); Analy Salles de Azevedo Melo, Arnaldo L. Colombo, Angela Satie Nishikaku (Universidade Federal de São Paulo, São Paulo, Brazil); Marijke Hendrickx, Dirk Stubbe (BCCM/IHEM, Scientific Institute of Public Health, Brussels, Belgium); Anne-Cécile Normand, Renaud Piarroux, Stéphane Ranque (Aix-Marseille University, Marseille, France); Françoise Dromer, Dea Garcia-Hermoso, Marie Desnos-Ollivier (Institut Pasteur, Paris, France); Michael Arabatzis, Aristea Velegraki (National Kapodistrian University of Athens, Athens, Greece); Gianluigi Cardinali (Università degli Studi di Perugia, Perugia, Italy); Laura Rosio Castañón, Maria Lucia Taylor, Conchita Toriello (Universidad Nacional Autónoma de México, Ciudad de México, México); Sybren de Hoog, Vincent Robert (Westerdijk Fungal Biodiversity Institute, Utrecht, The Netherlands); Célia Pais (University of Minho, Braga, Portugal); Wilhelm de Beer (FABI, Pretoria, South Africa); Marieka Gryzenhout (University of the Free State, Bloemfontein, South Africa); Josep Guarro, Jose F. Cano-Lira (Universitat Rovira i Virgili, Reus, Spain); Ariya Chindamporn (Chulalongkorn University, Bangkok, Thailand); Chompoonek Yurayart (Kasetsart University, Bangkok, Thailand); Barbara Robbertse, Conrad Schoch (NIH, Bethesda, MD, USA).

References

Alexander, B.D., and Pfaller, M.A. 2006. Contemporary tools for the diagnosis and management of invasive mycoses. Clin. Infect. Dis. **43**: S15–S27. doi:10.1086/504491.

Baires-Varguez, L., Cruz-García, A., Villa-Tanaka, L., Sánchez-García, S., Gaitán-Cepeda, L.A., Sánchez-Vargas, L.O., et al. 2007. Comparison of a randomly amplified polymorphic DNA (RAPD) analysis and ATB ID 32C system for identification of clinical isolates of different *Candida* species. Rev. Iberoam.

Published by NRC Research Press

Genome Downloaded from cdnsciencepub.com by University of Sydney on 05/03/21
For personal use only.

Genome Downloaded from cdnsciencepub.com by University of Sydney on 05/03/21
For personal use only.

Micol. **24**(2): 148–151. doi:10.1016/S1130-1406(07)70031-1. PMID: 17604435.

Balajee, S.A., Gribskov, J.L., Hanley, E., Nickle, D., and Marr, K.A. 2005. *Aspergillus lentulus* sp. nov., a new sibling species of *A. fumigatus*. Eukaryotic Cell, **4**(3): 625–632. doi:10.1128/ec.4.3.625-632.2005.

Balajee, S.A., Sigler, L., and Brandt, M.E. 2007. DNA and the classical way: identification of medically important molds in the 21st century. Med. Mycol. **45**(6): 475–490. doi:10.1080/13693780701449425. PMID:17710617.

Begerow, D., Nilsson, H., Unterseher, M., and Maier, W. 2010. Current state and perspectives of fungal DNA barcoding and rapid identification procedures. Appl. Microbiol. Biotechnol. **87**(1): 99–108. doi:10.1007/s00253-010-2585-4. PMID:20405123.

Beguin, H., Pyck, N., Hendrickx, M., Planard, C., Stubbe, D., and Detandt, M. 2012. The taxonomic status of *Trichophyton quinckeanum* and *T. interdigitale* revisited: a multigene phylogenetic approach. Med. Mycol. **50**(8): 871–882. doi:10.3109/13693786.2012.684153. PMID:22587727.

Bergman, A., Fernandez, V., Holmström, K.O., Claesson, B.E., and Enroth, H. 2007. Rapid identification of pathogenic yeast isolates by real-time PCR and two-dimensional melting-point analysis. Eur. J. Clin. Microbiol. Infect. Dis. **26**(11): 813–818. doi:10.1007/s10096-007-0369-2. PMID:17680284.

Brandt, M.E., Padhye, A.A., Mayer, L.W., and Holloway, B.P. 1998. Utility of random amplified polymorphic DNA PCR and TaqMan automated detection in molecular identification of *Aspergillus fumigatus*. J. Clin. Microbiol. **36**(7): 2057–2062. PMID:9650962.

Bruns, T.D., White, T.J., and Taylor, J.W. 1991. Fungal molecular systematics. Annu. Rev. Ecol. Syst. **22**(1): 525–564. doi:10.1146/annurev.es.22.110191.002521.

Ciardo, D.E., Schar, G., Bottger, E.C., Altwegg, M., and Bosshard, P.P. 2006. Internal transcribed spacer sequencing versus biochemical profiling for identification of medically important yeasts. J. Clin. Microbiol. **44**(1): 77–84. doi:10.1128/JCM.44.1.77-84.2006. PMID:16390952.

Crespo, A., Lumbsch, H.T., Mattsson, J.E., Blanco, O., Divakar, P.K., Articus, K., et al. 2007. Testing morphology-based hypotheses of phylogenetic relationships in Parmeliaceae (Ascomycota) using three ribosomal markers and the nuclear *RPB1* gene. Mol. Phylogenet. Evol. **44**(2): 812–824. doi:10.1016/j.ympev.2006.11.029. PMID:17276700.

Dendis, M., Horvath, R., Michalek, J., Ruzicka, F., Grijalva, M., Bartos, M., and Benedik, J. 2003. PCR-RFLP detection and species identification of fungal pathogens in patients with febrile neutropenia. Clin. Microbiol. Infect. **9**(12): 1191–1202. doi:10.1111/j.1469-0691.2003.00719.x. PMID:14686984.

Diguta, C.F., Vincent, B., Guilloux-Benatier, M., Alexandre, H., and Rousseaux, S. 2011. PCR ITS-RFLP: a useful method for identifying filamentous fungi isolates on grapes. Food Microbiol. **28**(6): 1145–1154. doi:10.1016/j.fm.2011.03.006. PMID: 21645813.

Fell, J.W., Boekhout, T., Fonseca, A., Scorzetti, G., and Statzell-Tallman, A. 2000. Biodiversity and systematics of basidiomycetous yeasts as determined by large-subunit rDNA D1/D2 domain sequence analysis. Int. J. Syst. Evol. Microbiol. **50**(3): 1351–1371. doi:10.1099/00207713-50-3-1351. PMID:10843082.

Ferrer, C., Colom, F., Frases, S., Mulet, E., Abad, J., and Alio, J. 2001. Detection and identification of fungal pathogens by PCR and by ITS2 and 5.8S ribosomal DNA typing in ocular infections. J. Clin. Microbiol. **39**: 2873–2879. doi:10.1128/JCM.39.8.2873-2879.2001. PMID:11474006.

Frisvad, J.C., and Samson, R.A. 2004. Polyphasic taxonomy of *Penicillium* subgenus *Penicillium*— a guide to identification of food and air-borne terverticillate penicillia and their mycotoxins. Stud. Mycol. **49**: 1–174.

Gardes, M., and Bruns, T.D. 1993. ITS primers with enhanced specificity for *Basidiomycetes* — application to the identification of mycorrhizae and rusts. Mol. Ecol. **2**(2): 113–118. doi:10.1111/j.1365-294X.1993.tb00005.x. PMID:8180733.

Gerrits van den Ende, A.H.G., and de Hoog, G.S. 1999. Variability and molecular diagnostics of the neurotropic species *Cladophialophora bantiana*. Stud. Mycol. **43**: 151–162.

Gilgado, F., Cano, J., Gené, J., and Guarro, J. 2005. Molecular phylogeny of the *Pseudallescheria boydii* species complex: proposal of two new species. J. Clin. Microbiol. **43**(10): 4930–4942. doi:10.1128/JCM.43.10.4930-4942.2005. PMID:16207945.

Hajibabaei, M., Singer, G.A., Hebert, P.D., and Hickey, D.A. 2007. DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. Trends Genet. **23**(4): 167–172. doi:10.1016/j.tig.2007.02.001. PMID:17316886.

Hebert, P.D., Cywinska, A., Ball, S.L., and deWaard, J.R. 2003*a*. Biological identifications through DNA barcodes. Proc. Biol. Sci. **270**(1512): 313–321. doi:10.1098/rspb.2002.2218. PMID: 12614582.

Hebert, P.D., Ratnasingham, S., and deWaard, J.R. 2003*b*. Barcoding animal life: cytochrome *c* oxidase subunit 1 divergences among closely related species. Proc. R. Soc. B Biol. Sci. **270**(S1): S96–S99. doi:10.1098/rsbl.2003.0025.

Hierro, N., González, A., Mas, A., and Guillamón, J.M. 2004. New PCR-based methods for yeast identification. J. Appl. Microbiol. **97**(4): 792–801. doi:10.1111/j.1365-2672.2004.02369.x. PMID: 15357729.

Hillis, D.M., and Dixon, M.T. 1991. Ribosomal DNA: molecular evolution and phylogenetic inference. Q. Rev. Biol. **66**(4): 411–453. doi:10.1086/417338. PMID:1784710.

Hofstetter, V., Miadlikowska, J., Kauff, F., and Lutzoni, F. 2007. Phylogenetic comparison of protein-coding versus ribosomal RNA-coding sequence data: a case study of the Lecanoromycetes (Ascomycota). Mol. Phylogenet. Evol. **44**(1): 412–426. doi:10.1016/j.ympev.2006.10.016. PMID:17207641.

Irinyi, L., Serena, C., Garcia-Hermoso, D., Arabatzis, M., Desnos-Ollivier, M., Vu, D., et al. 2015. International Society of Human and Animal Mycology (ISHAM)-ITS reference DNA barcoding database-the quality controlled standard tool for routine identification of human and animal pathogenic fungi. Med. Mycol. **53**(4): 313–337. doi:10.1093/mmy/myv008. PMID:25802363.

James, T.Y., Kauff, F., Schoch, C.L., Matheny, P.B., Hofstetter, V., Cox, C.J., et al. 2006. Reconstructing the early evolution of Fungi using a six-gene phylogeny. Nature, **443**(7113): 818–822. doi:10.1038/nature05110. PMID:17051209.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**(2): 111–120. doi:10.1007/BF01731581. PMID:7463489.

Kiss, L. 2012. Limits of nuclear ribosomal DNA internal transcribed spacer (ITS) sequences as species barcodes for Fungi. Proc. Natl. Acad. Sci. U.S.A. **109**(27): E1811. doi:10.1073/pnas.1207143109. PMID:22715287.

Klingspor, L., and Jalal, S. 2006. Molecular detection and identification of *Candida* and *Aspergillus* spp. from clinical samples using real-time PCR. Clin. Microbiol. Infect. **12**(8): 745–753. doi:10.1111/j.1469-0691.2006.01498.x. PMID:16842569.

Kulik, T., Fordoński, G., Pszczółkowska, A., Płodzień, K., and Łapiński, M. 2004. Development of PCR assay based on ITS2 rDNA polymorphism for the detection and differentiation of *sporotrichioides*. FEMS Microbiol. Lett. **239**(1): 181–186. doi:10.1016/j.femsle.2004.08.037. PMID:15451117.

Kurtzman, C., and Robnett, C. 1998. Identification and phylogeny of ascomycetous yeasts from analysis of nuclear large subunit (26S) ribosomal DNA partial sequences. Antonie van Leeuwenhoek, **73**(4): 331–371. doi:10.1023/A:1001761008817. PMID:9850420.

Leaw, S.N., Chang, H.C., Sun, H.F., Barton, R., Bouchara, J.-P.,

Genome Downloaded from cdnsciencepub.com by University of Sydney on 05/03/21
For personal use only.

and Chang, T.C. 2006. Identification of medically important yeast species by sequence analysis of the internal transcribed spacer regions. J. Clin. Microbiol. **44**(3): 693–699. doi:10.1128/JCM.44.3.693-699.2006. PMID:16517841.

Letourneau, A., Seena, S., Marvanová, L., and Bärlocher, F. 2010. Potential use of barcoding to identify aquatic hyphomycetes. Fungal Divers. **40**(1): 51–64. doi:10.1007/s13225-009-0006-8.

Librado, P., and Rozas, J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics, **25**(11): 1451–1452. doi:10.1093/bioinformatics/btp187. PMID:19346325.

Lieckfeldt, E., Meyer, W., and Börner, T. 1993. Rapid identification and differentiation of yeasts by DNA and PCR fingerprinting. J. Basic Microbiol. **33**(6): 413–425. doi:10.1002/jobm.3620330609. PMID:8271158.

Lindsley, M.D., Hurst, S.F., Iqbal, N.J., and Morrison, C.J. 2001. Rapid identification of dimorphic and yeast-like fungal pathogens using specific DNA probes. J. Clin. Microbiol. **39**(10): 3505–3511. doi:10.1128/JCM.39.10.3505-3511.2001. PMID:11574564.

Martin, C., Roberts, D., van Der Weide, M., Rossau, R., Jannes, G., Smith, T., and Maher, M. 2000. Development of a PCR-based line probe assay for identification of fungal pathogens. J. Clin. Microbiol. **38**(10): 3735–3742. PMID:11015393.

McLaughlin, D.J., Hibbett, D.S., Lutzoni, F., Spatafora, J.W., and Vilgalys, R. 2009. The search for the fungal tree of life. Trends Microbiol. **17**(11): 488–497. doi:10.1016/j.tim.2009.08.001.

Meyer, C.P., and Paulay, G. 2005. DNA barcoding: error rates based on comprehensive sampling. PLoS Biol. **3**(12): e422. doi:10.1371/journal.pbio.0030422. PMID:16336051.

Meyer, W., and Gams, W. 2003. Delimitation of *Umbelopsis* (*Mucorales*, *Umbelopsidaceae* fam. nov.) based on ITS sequence and RFLP data. Mycol. Res. **107**(3): 339–350. PMID:12825503.

Nilsson, R., Ryberg, M., Kristiansson, E., Abarenkov, K., Larsson, K., and Kõljalg, U. 2006. Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. PLoS ONE, **1**(1): e59. doi:10.1371/journal.pone.0000059. PMID:17183689.

Nilsson, R.H., Kristiansson, E., Ryberg, M., Hallenberg, N., and Larsson, K.H. 2008. Intraspecific ITS variability in the kingdom *Fungi* as expressed in the international sequence databases and its implications for molecular species identification. Evol. Bioinform. Online, **4**: 193–201. PMID:19204817.

Nilsson, R.H., Ryberg, M., Abarenkov, K., Sjokvist, E., and Kristiansson, E. 2009. The ITS region as a target for characterization of fungal communities using emerging sequencing technologies. FEMS Microbiol. Lett. **296**(1): 97–101. doi:10.1111/j.1574-6968.2009.01618.x. PMID:19459974.

O'Donnell, K. 1993. *Fusarium* and its near relatives. *In* The fungal holomorph: mitotic, meiotic and pleomorphic speciation in fungal systematics. *Edited by* D.R. Reynolds and J.W. Taylor. CAB International, Wallingford, U.K. pp. 225–233.

O'Donnell, K. 2000. Molecular phylogeny of the *Nectria haematococca-solani* species complex. Mycologia, **92**(5): 919–938. doi:10.2307/3761588.

O'Donnell, K., Sutton, D.A., Rinaldi, M.G., Sarver, B.A., Balajee, S.A., Schroers, H.J., et al. 2010. Internet-accessible DNA sequence database for identifying fusaria from human and animal infections. J. Clin. Microbiol. **48**(10): 3708–3718. doi:10.1128/JCM.00989-10. PMID:20686083.

O'Donnell, K., Rooney, A.P., Proctor, R.H., Brown, D.W., McCormick, S.P., Ward, T.J., et al. 2013. Phylogenetic analyses of *RPB1* and *RPB2* support a middle Cretaceous origin for a clade comprising all agriculturally and medically important fusaria. Fungal Genet. Biol. **52**: 20–31. doi:10.1016/j.fgb.2012.12.004. PMID:23357352.

Pryce, T.M., Palladino, S., Kay, I.D., and Coombs, G.W. 2003. Rapid identification of fungi by sequencing the ITS1 and ITS2 regions using an automated capillary electrophoresis system. Med. Mycol. **41**(5): 369–381. doi:10.1080/13693780310001600435. PMID:14653513.

Rehner, S.A., and Buckley, E. 2005. A *Beauveria* phylogeny inferred from nuclear ITS and EF1-a sequences: evidence for cryptic diversification and links to *Cordyceps* teleomorphs. Mycologia, **97**: 84–98. doi:10.1080/15572536.2006.11832842. PMID:16389960.

Robert, V., Szöke, S., Eberhardt, U., Cardinali, G., Seifert, K.A., Lévesque, C.A., et al. 2011. The quest for a general and reliable fungal DNA barcode. Open Appl. Inform. J. **5**(Suppl. 1-M6): 45–61. doi:10.2174/1874136301105010045.

Romanelli, A.M., Sutton, D.A., Thompson, E.H., Rinaldi, M.G., and Wickes, B.L. 2010. Sequence-based identification of filamentous basidiomycetous fungi from clinical specimens: a cautionary note. J. Clin. Microbiol. **48**(3): 741–752. doi:10.1128/JCM.01948-09.

Sandhu, G.S., Kline, B.C., Stockman, L., and Roberts, G.D. 1995. Molecular probes for diagnosis of fungal infections. J. Clin. Microbiol. **33**(11): 2913–2919. PMID:8576345.

Sangoi, A.R., Rogers, W.M., Longacre, T.A., Montoya, J.G., Baron, E.J., and Banaei, N. 2009. Challenges and pitfalls of morphologic identification of fungal infections in histologic and cytologic specimens: a ten-year retrospective review at a single institution. Am. J. Clin. Pathol. **131**(3): 364–375. doi:10.1309/AJCP99OOOZSNISCZ. PMID:19228642.

Schoch, C.L., Sung, G.H., Lopez-Giráldez, F., Townsend, J.P., Miadlikowska, J., Hofstetter, V., et al. 2009. The Ascomycota tree of life: a phylum-wide phylogeny clarifies the origin and evolution of fundamental reproductive and ecological traits. Syst. Biol. **58**(2): 224–239. doi:10.1093/sysbio/syp020. PMID:20525580.

Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A., and Chen, W. 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proc. Natl. Acad. Sci. U.S.A. **109**(16): 6241–6246. doi:10.1073/pnas.1117018109. PMID:22454494.

Scorzetti, G., Fell, J.W., Fonseca, A., and Statzell-Tallman, A. 2002. Systematics of basidiomycetous yeasts: a comparison of large subunit D1/D2 and internal transcribed spacer rDNA regions. FEMS Yeast Res. **2**(4): 495–517. doi:10.1111/j.1567-1364.2002.tb00117.x. PMID:12702266.

Seifert, K.A. 2009. Progress towards DNA barcoding of fungi. Mol. Ecol. Resour. **9**(S1): 83–89. doi:10.1111/j.1755-0998.2009.02635.x. PMID:21564968.

Seifert, K.A., and Samuels, G.J. 2000. How should we look at anamorphs? Stud. Mycol. **45**: 5–18.

Stielow, J.B., Lévesque, C.A., Seifert, K.A., Meyer, W., Iriny, L., Smits, D., et al. 2015. One fungus, which genes? Development and assessment of universal primers for potential secondary fungal DNA barcodes. Persoonia, **35**: 242–263. doi:10.3767/003158515X689135. PMID:26823635.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. **28**(10): 2731–2739. doi:10.1093/molbev/msr121. PMID:21546353.

Tanabe, Y., Watanabe, M.M., and Sugiyama, J. 2005. Evolutionary relationships among basal fungi (Chytridiomycota and Zygomycota): insights from molecular phylogenetics. J. Gen. Appl. Microbiol. **51**(5): 267–276. doi:10.2323/jgam.51.267. PMID:16314681.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**(22): 4673–4680. doi:10.1093/nar/22.22.4673. PMID:7984417.

Vilgalys, R., and Gonzalez, D. 1990. Organization of ribosomal

DNA in the basidiomycete *Thanatephorus praticola*. Curr. Genet. **18**(3): 277–280. doi:10.1007/BF00318394. PMID:2249259.

Vilgalys, R., and Hester, M. 1990. Rapid genetic identification and mapping of enzymatically amplified ribosomal DNA from several *Cryptococcus* species. J. Bacteriol. **172**(8): 4238–4246. doi:10.1128/jb.172.8.4238-4246.1990. PMID:2376561.

White, T.J., Bruns, T., Lee, S., and Taylor, J.W. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. *In* PCR protocols: a guide to methods and applications. *Edited by* M.A. Innis, D.H. Gelfand, J.J. Sninsky, and T.J. White. Academic Press, New York. pp. 315–322.

Genome Downloaded from cdnsciencepub.com by University of Sydney on 05/03/21
For personal use only.

Chapter 4
Primary and Secondary Fungal DNA Barcode Comparison

4.1 Dual DNA Barcoding for the Molecular Identification of the Agents of Invasive Fungal Infections

This manuscript was published in Frontiers in Microbiology in July 2019 (volume 10, article 1647, pages 1-13). Hoang M.T.V. was the primary author and performed the research, analysed the data, and prepared the manuscript. Irinyi L. was involved in sample processing, all molecular studies, and data analysis. Chen S.C. and Sorrell T. conceived the study. All members of the ISHAM working group for "Barcoding of Medical Fungi" performed the sample collection. Meyer W. was the corresponding author and conceived and coordinated the study and supervised the molecular and data analysis. All authors wrote and corrected the manuscript.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Minh Thuy Vi Hoang
30th June 2023

As a supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Professor Wieland Meyer
30th June 2023

frontiers in Microbiology

ORIGINAL RESEARCH
published: 18 July 2019
doi: 10.3389/fmicb.2019.01647

Dual DNA Barcoding for the Molecular Identification of the Agents of Invasive Fungal Infections

*Minh Thuy Vi Hoang[1,2], Laszlo Irinyi[1,2,3], Sharon C. A. Chen[1,3,4], Tania C. Sorrell[1,2,3], the ISHAM Barcoding of Medical Fungi Working Group and Wieland Meyer[1,2,3,5]**

[1] Molecular Mycology Research Laboratory, Centre for Infectious Diseases and Microbiology, Sydney Medical School, Westmead Clinical School, Faculty of Medicine and Health, The University of Sydney, Sydney, NSW, Australia, [2] The Westmead Institute for Medical Research, Westmead, NSW, Australia, [3] Marie Bashir Institute for Infectious Diseases and Biosecurity, The University of Sydney, Sydney, NSW, Australia, [4] Centre for Infectious Diseases and Microbiology Laboratory Services, Institute for Clinical Pathology and Medical Research, NSW Health Pathology, Westmead, NSW, Australia, [5] Research and Education Network, Westmead Hospital, Westmead, NSW, Australia

Invasive fungal infections, such as aspergillosis, candidiasis, and cryptococcosis, have significantly increased among immunocompromised people. To tackle these infections the first and most decisive step is the accurate identification of the causal pathogen. Routine identification of invasive fungal infections has progressed away from culture-dependent methods toward molecular techniques, including DNA barcoding, a highly efficient and widely used diagnostic technique. Fungal DNA barcoding previously relied on a single barcoding region, the internal transcribed spacer (ITS) region. However, this allowed only for 75% of all fungi to be correctly identified. As such, the *translational elongation factor 1α* (*TEF1α*) was recently introduced as the secondary barcode region to close the gap. Both loci together form the dual fungal DNA barcoding scheme. As a result, the ISHAM Barcoding Database has been expanded to include sequences for both barcoding regions to enable practical implementation of the dual barcoding scheme into clinical practice. The present study investigates the impact of the secondary barcode on the identification of clinically important fungal taxa, that have been demonstrated to cause severe invasive disease. Analysis of the barcoding regions was performed using barcoding gap analysis based on the genetic distances generated with the Kimura 2-parameter model. The secondary barcode demonstrated an improvement in identification for all taxa that were unidentifiable with the primary barcode, and when combined with the primary barcode ensured accurate identification for all taxa analyzed, making DNA barcoding an important, efficient and reliable addition to the diagnostic toolset of invasive fungal infections.

Keywords: identification, fungal DNA barcoding, dual barcoding system, internal transcribed spacer region, *translational elongation factor 1α*, ISHAM Barcoding Database, invasive fungal diseases

OPEN ACCESS

Edited by:
Saad J. Taj-Aldeen, Hamad Medical Corporation, Qatar

Reviewed by:
Zheng Wang, Yale University, United States
Huzefa A. Raja, The University of North Carolina at Greensboro, United States

***Correspondence:**
Wieland Meyer
wieland.meyer@sydney.edu.au

Specialty section:
This article was submitted to Fungi and Their Interactions, a section of the journal Frontiers in Microbiology

Received: *10 May 2019*
Accepted: *03 July 2019*
Published: *18 July 2019*

Citation:
Hoang MTV, Irinyi L, Chen SCA, Sorrell TC, the ISHAM Barcoding of Medical Fungi Working Group and Meyer W (2019) Dual DNA Barcoding for the Molecular Identification of the Agents of Invasive Fungal Infections. Front. Microbiol. 10:1647. doi: 10.3389/fmicb.2019.01647

INTRODUCTION

While AIDS-associated *Pneumocystis jirovecii* pneumonia (Pjp) and cryptococcosis have declined in developed countries due to widespread use of highly active antiretroviral treatment (Dromer et al., 2004; Morris et al., 2004; Rajasingham et al., 2017), the overall burden of invasive fungal diseases (IFDs), especially candidemia and invasive aspergillosis has increased worldwide

(Patterson, 2005; Maschmeyer, 2006; Pfaller et al., 2006b; Warnock, 2007; Pappas et al., 2010; Benedict et al., 2017). IFDs alone cause about 1.6 million deaths/year (Brown et al., 2012). The rise in incidence of IFDs is largely due to an increase in at-risk populations, especially immunocompromised individuals, such as recipients of solid organs or hematopoietic stem cell transplants, and patients with underlying chronic diseases (Kontoyiannis et al., 2010; Brown et al., 2012; Armstrong-James et al., 2014; Bitar et al., 2014; Schelenz et al., 2015). It is paramount that the management of invasive mycoses must be improved through advancements in prevention, diagnosis, treatment and surveillance (Denning, 2016; Kneale et al., 2016; Cole et al., 2017).

The majority of the current fungal diagnostic techniques are inadequate for the identification of all pathogenic fungi, which is a pre-requisite for timely initiation of appropriate antifungal therapy (Schelenz et al., 2015; Irinyi et al., 2016; Cole et al., 2017). Culture-based identification techniques rely on morphological and phenotypic characteristics, are often inaccurate, and lack in most cases species-specific features. Phenotypic traits also fail to differentiate between closely related species or species complexes with near-identical morphological characteristics but distinguishable genetic traits, such as the casual agents of many mold infections, including aspergillosis and scedosporiosis (Tavanti et al., 2005). Further, not all pathogenic fungi grow under laboratory conditions. Morphology based diagnostics are also time-consuming (7–14 days), highly laborious, and heavily dependent on the level of mycological expertise of the microscopist, making them unsuitable for rapid and reliable diagnosis (Irinyi et al., 2016).

Serological tests or fungal biomarkers, such as the *Aspergillus* antigen test or the *Cryptococcus* lateral flow assay (CrAg® LFA, IMMY, Norman, OK, United States), are available, but they are designed to identify specific pathogens (Chen et al., 2014; Schelenz et al., 2015; Cadena et al., 2016).

To overcome the limitations of standard phenotypic diagnosis and identification methods, culture- and "expert-free" methods capable of identifying fungi directly from biological specimens are needed. Sequence-based identification has proven to be more accurate than conventional methods in diagnostic clinical mycology (Ciardo et al., 2006; Balajee et al., 2007). Among the applied molecular techniques, DNA barcoding is one of the most promising and efficient methods, as it enables rapid identification of species and recognition of cryptic species across all fungal genera. As such, DNA barcoding has recently been established as the gold standard identification technique for fungal species and has been proven to be more accurate than conventional identification techniques (Ciardo et al., 2006; Balajee et al., 2007). Barcodes are standardized, easily amplified, universal short DNA sequences (500–800 bp), which are divergent at the species level enabling rapid identification by comparison with a validated reference sequence collection. To ensure consistency of identification, barcodes should be unique to a single species, and stable within each species (Hebert et al., 2003). Additionally, interspecies variation must exceed the intraspecies variation, generating a "break" in the distribution of distances, which is referred to as the "barcoding gap" (Meyer and Paulay, 2005).

DNA barcoding and its associated references databases [BOLD (Hebert et al., 2003), UNITE (Kõljalg et al., 2013), RefSeq at Genbank (Schoch et al., 2014), and the "ISHAM Barcoding Database" (Meyer et al., 2018)], plays a central role in the identification landscape, as it is the basis for all future methods, either culture dependent or culture independent (**Figure 1**).

After numerous candidate genetic loci were evaluated, with varying success rates, the internal transcribed spacer (ITS) region was established as the primary fungal DNA barcode by Schoch et al. (2012). The ITS region is composed of two non-coding and variable regions, ITS1 and ITS2, flanking the highly conserved 5.8S gene. They are located between the 18S [small subunit (SSU)] and 28S [large subunit (LSU)] genes in the nrDNA repeat (White et al., 1990). The advantage of the ITS region is, that it can be easily amplified from most fungal taxa, using universal primers, with the most commonly used ones being the ITS1, ITS2, ITS3, ITS4, and ITS5 (White et al., 1990). Fungal-specific primers were also designed to avoid cross reactivity with plant or animal DNA, such as SR6R and LR1 (Vilgalys and Hester, 1990), V9D, V9G, and LS266 (Gerrits van den Ende and de Hoog, 1999), IT2 (Beguin et al., 2012), ITS1F (Gardes and Bruns, 1993) and NL4b (O'Donnell, 1993; **Figure 2** and **Table 1**). To provide quality controlled reference primary fungal DNA barcode sequences the International Society for Human and Animal Mycology (ISHAM) "ITS DNA barcode database" was established in 2015 (Irinyi et al., 2015). The primary fungal DNA barcode region identifies up to 75% of the estimated ~700 pathogenic fungal species (de Hoog et al., 2014).

To address the shortcomings of the ITS region and hence close this identification gap, a secondary barcode was proposed in 2015 (Stielow et al., 2015). The *translational elongation factor 1α* (*TEF1α*) was selected due to its high species discrimination across fungal taxa and the ability to design universal primers, such as EF1-1018F (Al33F)/EF1-1620R (Al33R), EF1-1002F (Al34F)/EF1-1688R (Al34R) (Stielow et al., 2015), or EF1-983F/EF1-1567R (Rehner and Buckley, 2005; **Figure 2** and **Table 1**). To accommodate the secondary fungal DNA barcode a dedicated database that would eventually include all medically relevant fungal species was established to complement the ISHAM-ITS database, which contains only quality-controlled *TEF1α* sequences obtained from taxonomically verified fungal cultures (Meyer et al., 2018). Both databases have been combined within the "ISHAM Barcoding Database",[1] which was launched in November 2017 at the 7th International DNA barcoding conference at the Kruger National Park in South Africa. The combined dual barcode database provides the medical and veterinary community with quality-controlled primary and secondary fungal DNA barcodes reference sequences. This database is publicly available, and its quality-controlled sequences are shared with other major databases, including RefSeq within GenBank (Benson et al., 2014) and UNITE (Kõljalg et al., 2013).

To date, the clinical utility and accuracy of the dual DNA barcoding system for identification of pathogenic fungi has not been assessed. This study aimed to increase the number of reference secondary barcode sequences for

[1] http://its.mycologylab.org/

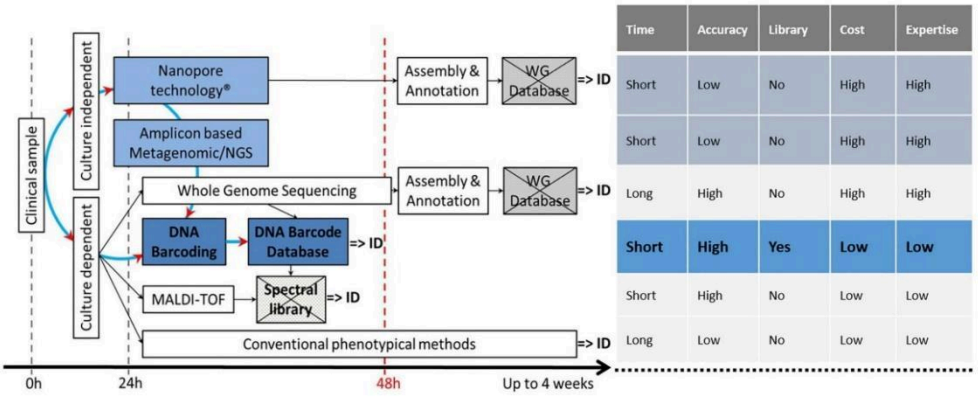


FIGURE 1 | Current and potentially available techniques for fungal ID, showing the central role of the reference DNA barcode database and the pros and cons of all techniques. Crossed boxes indicate missing databases. DNA barcoded reference strains are fundamental to build up a spectral library.

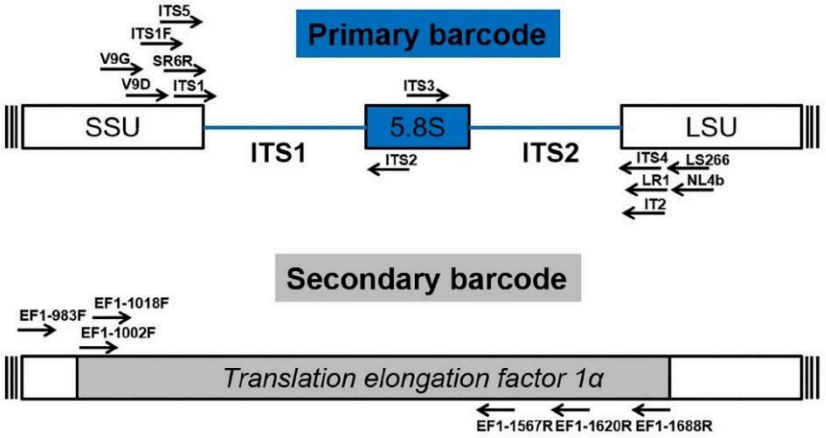


FIGURE 2 | Schematic structure of the primary (ITS) and secondary (*TEF1α*) fungal DNA barcode regions indicating universal primers for their amplification.

pathogenic fungal species and to compare the accuracy and resolution of the primary and secondary fungal DNA barcodes separately or in combination, focusing on fungi causing invasive fungal infections.

MATERIALS AND METHODS

Cultures

To generate quality controlled *TEF1α* sequences 270 strains, representing 90 human/animal pathogenic fungal species, were used (see **Supplementary Table 1**).

DNA Extraction

DNA was isolated and purified from cultures using either previously described manual method (Ferrer et al., 2001) or the Quick-DNA Fungal/Bacterial Kit (D6007, Zymo Research) according to the manufacturer's instructions.

DNA Barcode Generation

The secondary fungal DNA barcoding region (*TEF1α*) was amplified using the primers described by Stielow et al. (2015), including: EF1-1018F (Al33F) (5′ GAYTTCATCAAGAACATGAT 3′) and EF1-1620R (Al33R) (5′ GACGTTGAADCCRACRTTGTC 3′) being used together and

TABLE 1 | Universal and fungal specific primers for the dual DNA barcoding scheme.

Barcode	Primer name	Sequence	References
Primary barcode	ITS1	5′ TCCGTAGGTGAACCTGCGG 3′	White et al., 1990
	ITS2	5′ GCTGCGTTCTTCATCGATGC 3′	White et al., 1990
	ITS3	5′ GCATCGATGAAGAACGCAGC 3′	White et al., 1990
	ITS4	5′ TCCTCCGCTTATTGATATGC 3′	White et al., 1990
	ITS5	5′ GGAAGTAAAAGTCGTAACAAGG 3′	White et al., 1990
	ITS1F	5′ CTTGGTCATTTAGAGGAAGTAA 3′	Gardes and Bruns, 1993
	IT2	5′ CCTCCGCTTATTGATATGCTTAGG 3′	Beguin et al., 2012
	SR6R	5′ AAGTATAAGTCGTAACAAGG 3′	Vilgalys and Hester, 1990
	LR1	5′ GGTTGGTTTCTTTCCT 3′	Vilgalys and Hester, 1990
	V9D	5′ TTAAGTCCCTGCCCTTTGTA 3′	Gerrits van den Ende and de Hoog, 1999
	V9G	5′ TACGTCCCTGCCCTTTGTA 3′	Gerrits van den Ende and de Hoog, 1999
	LS266	5′ GCATTCCCAAACAACTCGACTC 3′	Gerrits van den Ende and de Hoog, 1999
	NL4b	5′ GGATTCTCACCCTCTATGAC 3′	O'Donnell, 1993
Secondary barcode	EF1-1002F	5′ TTCATCAAGAACATGAT 3′	Stielow et al., 2015
	EF1-1018F	5′ GAYTTCATCAAGAACATGAT 3′	Stielow et al., 2015
	EF1-1620R	5′ GACGTTGAADCCRACRTTGTC 3′	Stielow et al., 2015
	EF1-1688R	5′ GCTATCATCACAATGGACGTTCTTGGAG 3′	Stielow et al., 2015
	EF1-983F	5′ GCYCCYGGHCAYCGTGAYTTYAT 3′	Rehner and Buckley, 2005
	EF1-1567R	5′ ACHGTRCCRATACCACCRATCTT 3′	Rehner and Buckley, 2005

EF1-1002F (Al34F) (5′ TTC ATCAAGAACATGAT 3′) and EF1-1688R (Al34R) (5′ CTATCATCACAATGG ACGTTCTTGGAG 3′) being used together (Stielow et al., 2015). The primer set Al33F-Al33R was first used to amplify the *TEF1α* region. If amplification was unsuccessful then it was repeated using the second primer set Al34F-Al34R. Both primer pairs used the following PCR amplification protocol: 5 min initial denaturing at 94°C, followed by 40 of 50 s at 94°C, 50 s annealing at 48°C, 50 s at 72°C and 7 min final extension at 72°C (Stielow et al., 2015). Amplification success was visualized through gel electrophoresis in a 1.5% agarose gel containing ethidium bromide (EtBr) with ultra-violet illumination. Successfully amplified PCR products were sent for commercial sequencing, e.g., Macrogen Inc., South Korea in both forward and reverse directions. Bidirectional sequenced were assembled and edited using Sequencher® ver. 5.3. (Gene Codes Corporation, Ann Arbor, MI, United States). Sequences were manually checked to resolve ambiguous bases on the forward and reverse trace files considering the PHRED scores received.

DNA Barcode Analysis

The sequences for each taxon were aligned with the program CLUSTALW (Thompson et al., 1994) part of the software MEGA ver. 7 (Larkin et al., 2007). Resulting alignments were checked visually and edited when needed.

The intraspecies diversity was estimated by calculating the average nucleotide diversity (π) within species, where there were sequences available from more than three strains. The proportion of nucleotide differences in all haplotypes in the sample was derived using the software DnaSP ver. 5.10.01 (Librado and Rozas, 2009).

Individual analysis of the primary and secondary barcoding regions was performed through barcoding gap analysis. Additionally, analysis was performed on the combined barcoding regions. Taxonomic groups were selected for analysis if represented by three or more species, and if those species were represented by three or more strains. Furthermore, taxonomic groups were required to include at least one species that has been demonstrated to cause invasive fungal infections. Only strains with both primary and secondary barcodes in the database where included in the analysis. When available, more specific taxonomic groups such as clades were selected over genera. Sequences for each genus were aligned and cut to equal length using CLUSTALW as present in MEGA ver. 7 (Larkin et al., 2007; Kumar et al., 2016). Genetic distances between each strain were calculated using the Kimura 2-parameter model (K2P) and the intraspecies and interspecies genetic distances were compared (Kimura, 1980). Intra- and interspecies genetic distances were graphed against frequency and analyzed for barcoding gaps. Barcoding gaps were defined by the presence of a distinct difference between the largest intraspecies genetic distance and smallest interspecies genetic distance (Irinyi et al., 2015), i.e., there was no common x value between the intra- and interspecies groups.

The introduction of the secondary fungal DNA barcode was accessed for each genus on the basis that there would be an improvement in DNA barcoding if the secondary barcode generated a barcoding gap when the primary barcode did not, or if the overlap between the intra- and interspecies genetic distances was reduced.

RESULTS

The study produced 270 new quality controlled secondary fungal barcode sequences, covering 90 pathogenic fungal species

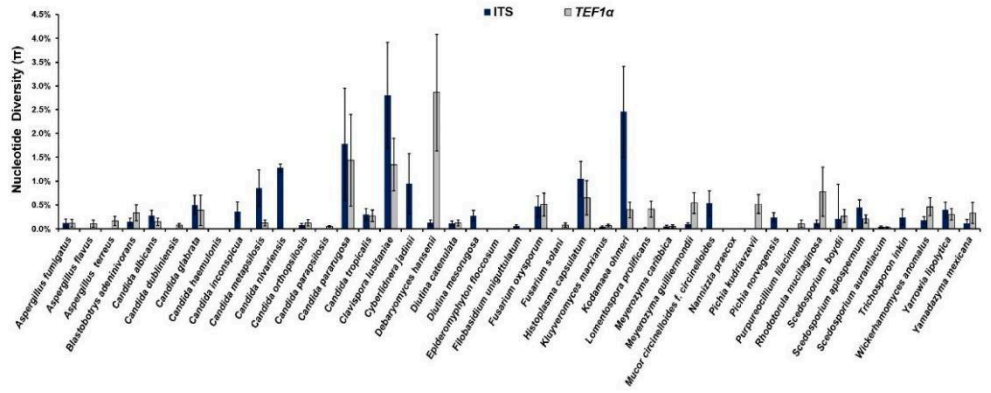


FIGURE 3 | Intraspecies variation for species, which are represented by more than three strains, in the internal transcribed spacer (ITS) (blue bars) primary fungal DNA barcode compared with the *translation elongation factor 1α* (*TEF1α*) (gray bars) secondary fungal DNA barcode.

(**Supplementary Table 1**). Overall, the PCR success rate was high within the 270 *TEF1* sequences. 220 secondary barcodes were generated using the Al33F–Al33R primer pair and 50 were amplified with the Al34F–Al34R primer pair. There was no trend in amplification success in different fungal species. However, the Al34F–Al34R primer set was required to amplify all strains of *Aspergillus niger*, *Candida albicans*, *Candida dubliniensis*, *Kluyveromyes marxianus*, and *Pichia kudriavzevii*. There was unsuccessful amplification with both primer sets for some strains of *Cladosporium* spp., *Rhodotorula* spp., and *Trichosporon* spp.

All sequences were submitted to the ISHAM Barcoding Database (see footnote 1). The length of the ITS and partial *TEF1α* sequences in the database ranges 285–791 and 534–1002 bp, respectively.

The analysis of the nucleotide diversity (π) of 43 fungal species with more than three strains in the ISHAM barcoding database showed that the *TEF1α* region is less diverse than the ITS region in most species (**Figure 3**). The intraspecies variation of *TEF1α* was for most species below 1.5%, confirming the secondary barcode as a more discriminator marker. According to the selection criteria four different taxonomic groups were selected as proof of principle for the dual barcoding system. These included the two genera, *Diutina* and *Scedosporium*, and the two taxonomic clades, *Lodderomyces* and *Pichia*. The *Diutina* genus was selected for analysis despite one of the species, *Diutina rugosa*, being represented by only two strains as this genus is newly established that causes rare disease (**Table 2**). The species and number of strains included in these analyses are outlined in **Table 2**.

Diutina Species

Barcoding gap analysis of the genus *Diutina* demonstrated no overlap between the intraspecies and interspecies genetic distances with either the primary or secondary fungal DNA barcodes (**Figure 4**). There was similarly no overlapping region for the barcoding gap analysis of the combination of the primary and secondary barcodes (**Figure 4**). As such, both fungal barcodes and the combination of the barcodes generated appropriate barcoding gaps.

The *Lodderomyces* Clade

For the *Lodderomyces* clade, the intraspecies and interspecies genetic distances, generated by the primary fungal barcode produced an overlapping region thereby indicating that there was no barcoding gap (**Figure 4**). The barcoding gap analysis for the secondary DNA barcode (*TEF1α*) demonstrated no overlap in the intraspecies and interspecies genetic distances for *Lodderomyces* thus showing a clear barcoding gap (**Figure 4**). The *Lodderomyces* clade analysis using the combination of both the ITS and *TEF1α*

TABLE 2 | Number of fungal strains with primary and secondary barcodes present in the ISHAM Barcoding Database analyzed in this study.

Taxonomic group	Species/Species complex	Number of strains
Diutina	*Diutina catenulata*	11
	Diutina mesorugosa	9
	Diutina rugosa	2
Lodderomyces clade	*Candida albicans*	13
	Candida dubliniensis	12
	Candida metapsilosis	7
	Candida orthopsilosis	11
	Candida parapsilosis	26
	Candida tropicalis	16
Pichia clade	*Candida inconspicua*	4
	Pichia kudriavzevii	15
	Pichia norvegensis	10
Scedosporium	*Scedosporium apiospermum*	6
	Scedosporium aurantiacum	13
	Scedosporium boydii	4

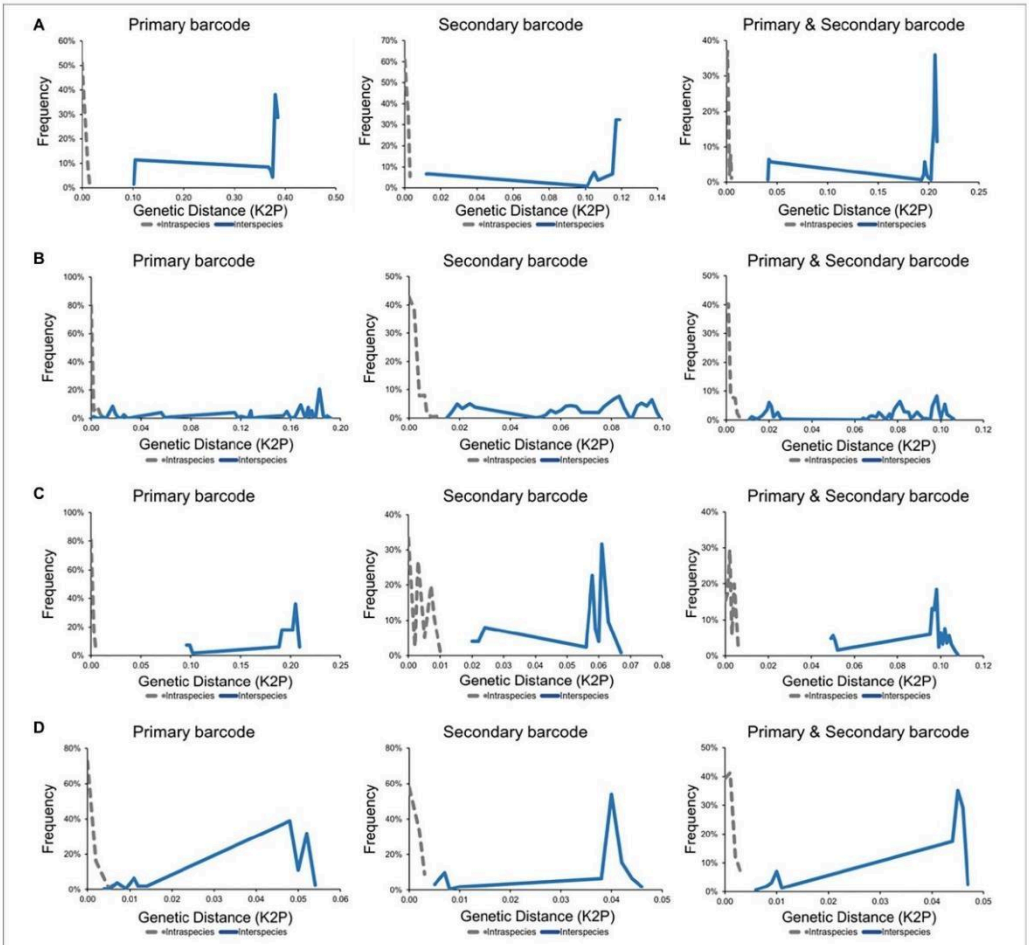


FIGURE 4 | Barcoding gap analyses for fungal taxa including causative agents of invasive fungal disease with intraspecies (gray) and interspecies (blue) genetic distances calculated using the Kimura 2-parameter model. **(A)** *Diutina* genus (*Diutina catenulata*, *Diutina mesorugosa*, and *Diutina rugosa*). **(B)** *Lodderomyces*-clade (*Candida albicans*, *Candida dublinsiensis*, *Candida metapsilosis*, *Candida orthopsilosis*, *Candida parapsilosis*, and *Candida tropicalis*). **(C)** *Pichia*-clade (*Candida inconspicua*, *Pichia kudriavzevii*, and *Pichia norvegensis*). **(D)** *Scedosporium* genus (*Scedosporium apiospermum*, *Scedosporium aurantiacum*, and *Scedosporium boydii*).

regions resulted in the generation of a barcoding gap being generated (**Figure 4**).

The *Pichia* Clade

Barcoding gap analysis of the *Pichia* clade did not reveal any overlapping regions in the genetic distances for either the ITS or *TEF1*α regions (**Figure 4**). As there were no overlapping regions this demonstrated that there are barcoding gaps for both the primary and secondary barcode. These results were similar when both barcoding regions were combined as there was a barcoding gap produced (**Figure 4**).

Scedosporium Species

Barcoding gap analysis of the primary fungal DNA barcode showed an overlapping region (**Figure 4**). For the secondary barcode, the barcoding gap analysis revealed that there was no overlap between the intraspecies and interspecies genetic distances and so a barcoding gap was present

(**Figure 4**). The barcoding gap analysis of the combination of both barcodes, produced a barcoding gap as there was no overlap between the intraspecies and interspecies genetic distances (**Figure 4**).

DISCUSSION

ISHAM Barcoding Database Expansion

Accurate and rapid routine diagnostic tools are essential to reduce the burden of fungal disease. DNA barcoding requires reliable quality-controlled reference sequences for its implementation. As such, fungal DNA Barcoding databases need to have high taxonomic coverage and reliable sequences to ensure users can accurately match sample sequences to the correct reference sequences. Additionally, the more sequences from different strains for each species are represented, the higher is the accuracy of the identification, as this reflects more realistically the existing species variation.

In this study, we generated 270 secondary barcodes, which could be amplified reliably using either the primer set Al33F–Al33R or the primer set Al34f–A134R (**Figure 2**), with a few exceptions where no amplifications where obtained for some strains of *Cladosporium* spp., *Rhodotorula* spp., or *Trichosporon* spp. In those cases, modification of the amplification conditions, such as touchdown PCR, may be appropriate.

As a result of this study the ISHAM Barcoding Database contains, as of the May 10, 2019, 4091 primary (ITS) and 868 secondary barcode (*TEF1α*) sequences covering 640 and 129 pathogenic fungal species, respectively. From those 2681 ITS and 613 *TEF1α* sequences representing 252 species cause invasive fungal diseases. There are an estimated 700 pathogenic fungal species, with the majority rarely causing infections. As such, the taxonomic coverage of ITS sequences within the "ISHAM Barcoding Database" is close to covering all pathogenic fungal species, whilst the *TEF1α* sequences cover less than one fifth. In addition, 59 species are currently only represented by one *TEF1α* sequence, which limits identification due to underrepresented intraspecies variation. As the "ISHAM Barcoding Database" has only been recently expanded to include *TEF1α* sequences, it is expected that as more *TEF1α* sequences are submitted, these issues will be resolved. Currently the "ISHAM Barcoding Database" has enabled pragmatic implementation of the dual barcode scheme for the identification of pathogenic fungi.

Comparison of the Primary and Secondary Barcodes

Since the establishment of the secondary fungal DNA barcode the improvement of the identification accuracy for pathogenic fungi has not been assessed. In this study, we compared for the first time the primary and secondary fungal DNA barcodes via barcoding gap analysis based on intraspecies and interspecies genetic distance values as calculated using the K2P model (Kimura, 1980). In addition, the combination of both barcodes was tested and compared to the individual barcodes.

Diutina Species (Previously Belonging to the Genus *Candida*)

The genus *Diutina* was established only recently. This group of fungal species was previously part of the genus *Candida* (Kurtzman et al., 2011; Khunnamwong et al., 2015). *Diutina* species are uncommon agents of disease, but are associated with nosocomial infections, with fatality rates of up to 70% (Padovan et al., 2013). *Diutina catenulata* and *Diutina rugosa* are well known causes of fungemia in immunocompromised patients and the seriously ill (Radosavljevic et al., 1999; Minces et al., 2009; Behera et al., 2010; Ha et al., 2018). *D. rugosa* is notable for its increasing resistance against multiple antifungal agents worldwide, thereby increasing the need for early detection and identification to enable timely initiation of effective treatment (Pfaller et al., 2006a).

An exception was made for the genus *Diutina* to be included into this analysis, as *D. rugosa* is only represented by two strains in the analysis (**Table 2**). The inclusion of the species was important as *Diutina mesorugosa* and *D. rugosa* have been proposed as synonyms with the variability of ITS sequences used as supporting evidence (Ming et al., 2019).

Barcoding gap analysis indicates that these species can be clearly identified as separate species. The *Diutina* species are representative of the 75% of pathogenic fungi that are accurately identified by the primary barcode (Irinyi et al., 2016). The secondary barcode and the combination of both barcodes equally identified these species (**Figure 4**), demonstrating the same level of resolution as the primary barcode (Stielow et al., 2015).

The *Lodderomyces* Clade

The *Lodderomyces* clade contains 21 *Candida* species, including some of the major pathogenic fungal species (Kurtzman et al., 2011). This clade contains three of the five most common pathogenic *Candida* species: *Candida albicans*, *Candida dubliniensis*, and *Candida parapsilosis* (Spampinato and Leonardi, 2013). These species cause a wide spectrum of diseases, from cutaneous infections to fatal disseminated septicemia (Tavanti et al., 2005). *Candida* species are predominantly commensals in the body and as such these infections largely target immunocompromised patients and are commonly the cause of nosocomial outbreaks (Brown et al., 2012).

Candida is a polyphyletic genus with a highly complex taxonomic history (Kurtzman et al., 2011). In the latest edition of The Yeasts A Taxonomic Study, 314 different *Candida* species were described excluding species with anamorph-teleomorph linkages (Kurtzman et al., 2011). The genus *Candida* was initially intended to include undifferentiated yeasts that cannot be identified by phenotype (Berkhout, 1923). As such there is little stability in the genus with some species being linked to others as anamorphs and the introduction of newly found species. With the advanced use of genomics in taxonomic studies the composition of species and the *Lodderomyces* clade are expected to change.

The barcoding gap analysis of the primary barcode was unable to generate a barcoding gap and so was unable to accurately identify all species of the clade (**Figure 4**).

Upon closer inspection, the lack of a barcoding gap was due to the close relationships between *Candida orthopsilosis* and *Candida parapsilosis,* which form part of the *Candida parapsilosis* species complex. This species complex was previously representing a single species, *Candida parapsilosis,* with three subgroups that were then revised to be separate species based on various molecular techniques, including sequencing of the ITS region. The barcoding gap analysis of this study, however, did not reflect this differentiation and so ITS was unable to identify these different species. Barcoding gap analysis of the secondary barcode did produce a barcoding gap and thereby demonstrated that *TEF1α* could accurately identify all species of the Lodderomyces clade (**Figure 4**). As such, the introduction of the dual barcoding scheme resulted in an accurate identification of these pathogenic fungal species.

The *Pichia* Clade

The *Pichia* clade is composed of 20 species. In our study, *Candida inconspicua* was also included in the analysis, as it is thought to be the anamorph for *Pichia cactophila* (Kurtzman and Robnett, 1998; Kurtzman et al., 2008, 2011). *Pichia kudriavzevii* and *Pichia norvegensis* are predominantly linked to nosocomial pathogens and may cause outbreaks of infections (Nagarathnamma et al., 2017). *Candida inconspicua* has also been found in clinical samples whilst the other species of the *Pichia* clade have not yet been found to be medically relevant (Kurtzman et al., 2011; Guitard et al., 2013). These species also cause invasive infections in patients with a highly compromised immune systems, and are often found under previous names in the literature (Kurtzman et al., 2011; Guitard et al., 2013; Schuster et al., 2013; Sanclemente et al., 2015; Douglass et al., 2018).

Similarly, to the genus *Diutina*, barcoding gap analysis of the *Pichia* clade revealed barcoding gaps for the primary barcode, secondary barcode and the combination of both barcodes (**Figure 4**), resulting in an accurate identification of all species of the *Pichia* clade.

Scedosporium Species

Fungi belonging to the genus *Scedosporium* cause a wide variety of diseases from localized infections to invasive diseases (Cortez et al., 2008). *Scedosporium* species mainly cause infections in patients with a compromised immune systems largely due to underlying illnesses such as solid organ transplantation, cystic fibrosis, leukemia and bone marrow transplantation (Tamm et al., 2001; Allen et al., 2013; Kubisiak-Rzepczyk et al., 2013; Yu et al., 2013; Johnson et al., 2014). *Scedosporium* spp. can also cause infection in immunocompetent patients (Ceccarelli et al., 2012; Agatha et al., 2014; Cruysmans et al., 2015).

The taxonomy of the genus *Scedosporium* is complex and has changed multiple times over the last decade. Previously, the nomenclature of the genus ruled *Scedosporium* to be the anamorph name whilst *Pseudallescheria* and *Petriella* were the teleomorph names (Rainer and de Hoog, 2006; Cortez et al., 2008; Lu et al., 2011). The two major clinical species were *Scedosporium prolificans* and *Scedosporium apiospermum* and its teleomorph form, *Pseudallescheria boydii* (Cortez et al., 2008). Multilocus sequencing, morphological analysis and physiological testing later demonstrated that *S. apiospermum* and *P. boydii* are separate species and not anamorphs teleomorph pairs (Gilgado et al., 2008). In 2011, the new nomenclature rules dictating "one fugus = one name" led to the fact that *Scedosporium* took precedence over *Pseudallescheria* as the genus name (Hawksworth, 2011; Lackner et al., 2014). The composition of the genus was also changed with *S. prolificans* being removed from the genus and renamed to *Lomentospora prolificans* (Lackner et al., 2014). The separation between *S. apiospermum* and *S. boydii* was reinforced, however, it was noted that there was no clinical difference between the species, as such all species are referred to the *S. apiospermum* species complex (Lackner et al., 2014).

In this study, there was no barcoding gap present for the primary barcode indicating that *Scedosporium* species cannot be accurately identified using this region alone (**Figure 4**). The predominant reason for the overlap in intraspecies and interspecies genetic distances was the high similarity between *S. apiospermum* and *S. boydii*. This was reflective of the nomenclature history as these species were previously thought to be an anamorph – teleomorph pair. Barcoding gap analysis of the secondary barcoding region introduced a barcoding gap and hence allowed for the accurate identification of all *Scedosporium* species (**Figure 4**). As the separation of *S. apiospermum* and *S. boydii* was established via multilocus sequence analysis, it is reassuring that *TEF1α* can resolve all *Scedosporium* species to the same degree (Gilgado et al., 2008). With the introduction of the dual barcoding scheme, all *Scedosporium* species can now be accurately identified.

Assessment of the Dual Barcoding Scheme

Meaningful implementation of the dual DNA barcoding scheme depends on demonstration of an improvement in the identification accuracy of fungal species causing infections when adding the secondary barcode, *TEF1α*. To demonstrate this the two fungal barcodes were analyzed separately and in combination from selected species, which had sequences from both barcodes being in the "ISHAM Barcoding Database," using barcoding gap analysis through the generation of genetic distances by the K2P model (Kimura, 1980).

Although K2P has been used in all DNA barcoding studies since its introduction by Hebert et al. (2003), the accuracy of this model has been questioned (Kimura, 1980; Hebert et al., 2003). In the selection of ITS region as the primary fungal DNA barcode, the uncorrected p-distances were used for barcoding gap analysis whereas in the case of the secondary fungal DNA barcode (*TEF1α*) analysis the K2P method was applied (Schoch et al., 2012; Stielow et al., 2015). The K2P region was selected to generate barcoding gaps over uncorrected p-distances as K2P has been found to generate larger barcoding gaps for most data sets (Srivathsan and Meier, 2012). Additionally, in datasets where uncorrected p-distances were preferred, the

performance of K2P was found to be similar (Srivathsan and Meier, 2012). More complicated models have been applied in some studies in an attempt to improve the accuracy of the K2P model, however, the differences were minimal, prompting us to use K2P in our study (Collins et al., 2012; Barley and Thomson, 2016).

Comparison of the number of pathogenic fungal species and strains represented in the "ISHAM Barcoding Database" highlighted the deficit in the number of *TEF1α* reference sequences. Therefore, the value of this study is limited as only strains with both barcodes available were used for the analysis. More species represented by more sequences would give a more exact assessment about the value and resolution power of the dual barcoding concept.

DNA barcode gab analysis showed that two of the groups (*Diutina* and *Pichia*) tested were accurately identified using the primary barcode and the secondary barcode on its own, demonstrating that both barcodes had similar resolution. For the remaining two groups (*Lodderomyces* and *Scedosporium*) the primary barcode was not able to produce barcoding gaps whilst the secondary did, demonstrating that the secondary barcode had a higher resolution power. In addition, the combination of both barcodes increased the discriminatory power enabling a more accurate identification. These results indicate, that the application of the dual barcoding system drastically improves species identification in cases where a single barcoding system is unable to do so.

Implementation of the Dual Barcoding Scheme

We envisage that the proposed dual barcoding scheme can be applied in the routine diagnostic setting in a stepwise procedure. After a clinical specimen is obtained, the unknown fungal isolate is first assessed based on its morphologic and/or biochemical characteristics. Then those unknown fungal isolates which lack obvious morphological characteristics or result in unclear biochemical profiles should be subjected to DNA isolation and primary fungal DNA barcoding (ITS1/2 region). If the obtained sequence shows less than 98.5% identity to a given ITS reference sequence in the database, the secondary fungal DNA barcode (*TEF1α*) should be obtained to achieve a final species identification (**Figure 5**).

The dual fungal DNA barcoding scheme is an efficient diagnostic tool for the identification of agents of invasive mycoses and can be confidently implemented into routine

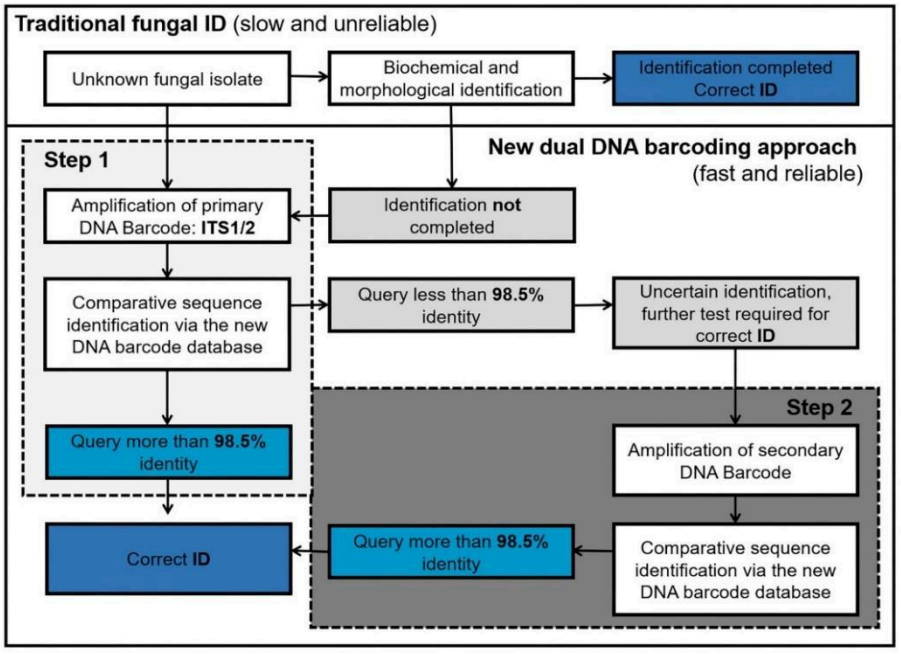


FIGURE 5 | Workflow of DNA barcoding based fungal ID (Steps 1 and 2 can be done in parallel to speed up the process).

diagnostics. Its implementation along with the expansion of the "ISHAM Barcoding Database" will lead to a paradigm shift in fungal disease diagnostics, enabling a highly accurate identification of the agents of fungal disease. The timely initiation of appropriate therapy will improve patient outcomes through the reduction in morbidity and mortality, prevent the use of unnecessary or inappropriate antifungal drugs, minimizing drug toxicity and resistance development and greatly reduce the associated health care cost.

CALL FOR DATA SUBMISSION

To achieve a 100% coverage of all pathogenic fungi and reflect the intraspecies variation for both the primary and secondary fungal DNA barcode we call for the submission of quality-controlled reference sequences of the ITS region and the *TEF1α* for inclusion into the "ISHAM Barcode Database." Please contact the curators of the database WM at wieland.meyer@sydney.edu.au and LI at laszlo.irinyi@sydney.edu.au.

MEMBERS OF THE ISHAM BARCODING OF MEDICAL FUNGI WORKING GROUP

Michael Arabatzis, National and Kapodistrian University of Athens, Athens, Greece; Ian Arthur, PathWest Laboratory Medicine, Nedlands, WA, Australia; Jose F. Cano-Lira, Universitat Rovira i Virgili, Reus, Spain; Gianluigi Cardinali, Università degli Studi di Perugia, Perugia, Italy; Laura Rosio Castañón, Universidad Nacional Autónoma de México, Mexico City, Mexico; Sharon Chen, Westmead Hospital, University of Sydney, Sydney, NSW, Australia; Wen Chen, Agriculture and Agri-Food Canada, Ottawa, ON, Canada; Ariya Chindamporn, Chulalongkorn University, Bangkok, Thailand; Arnaldo L. Colombo, Universidade Federal de São Paulo, São Paulo, Brazil; Marie Desnos-Ollivier, Institut Pasteur, Paris, France; Wilhelm de Beer, FABI, Pretoria, South Africa; Sybren de Hoog, Westerdijk Fungal Biodiversity Institute, Utrecht, Netherlands; Françoise Dromer, Institut Pasteur, Paris, France; Dea Garcia-Hermoso, Institut Pasteur, Paris, France; Marieka Gryzenhout, University of the Free State, Bloemfontein, South Africa; Josep Guarro, Universitat Rovira i Virgili, Reus, Spain; Catriona Halliday, University of Sydney/Westmead Hospital, Sydney, NSW, Australia; Marijke Hendrickx, BCCM/IHEM, Scientific Institute of Public Health, Brussels, Belgium; Sabine Huhndorf, The Field Museum, Chicago, IL, United States; Laszlo Irinyi, Westmead Hospital, University of Sydney, Sydney, NSW, Australia; C. Andre Levesque, Agriculture and Agri-Food Canada, Ottawa, ON, Canada; Wieland Meyer, Westmead Hospital, University of Sydney, Sydney, NSW, Australia; Maria Luiza Moretti, University of Campinas, Campinas, Brazil; Mauro de Medeiros Muniz, Instituto de Pesquisa Clínica Evandro Chagas (IPEC), Fundação Oswaldo Cruz (FIOCRUZ), Rio de Janeiro, Brazil; Analy Salles de Azevedo Melo, Universidade Federal de São Paulo, São Paulo, Brazil; Angela Satie Nishikaku, Universidade Federal de São Paulo, São Paulo, Brazil; Anne-Cécile Normand, Aix-Marseille University, Marseille, France; Célia Pais, University of Minho, Braga, Portugal; Renaud Piarroux, Aix-Marseille University, Marseille, France; Stéphane Ranque, Aix-Marseille University, Marseille, France; Barbara Robbertse, National Center for Biotechnology Information, Bethesda, MD, United States; Vincent Robert, Westerdijk Fungal Biodiversity Institute, Utrecht, Netherlands; Conrad L. Schoch, National Center for Biotechnology Information, Bethesda, MD, United States; Keith A. Seifert, Agriculture and Agri-Food Canada, Ottawa, ON, Canada; Célia Maria de Almeida Soares, Universidade Federal de Goiás, Goiânia, Brazil; Tania C. Sorrell, Westmead Hospital, University of Sydney, Sydney, NSW, Australia; John L. Spouge, National Center for Biotechnology Information, Bethesda, MD, United States; Dirk Stubbe, BCCM/IHEM, Scientific Institute of Public Health, Brussels, Belgium; Maria Lucia Taylor, Universidad Nacional Autónoma de México, Mexico City, Mexico; Conchita Toriello, Universidad Nacional Autónoma de México, Mexico City, Mexico; Aristea Velegraki, National and Kapodistrian University of Athens, Athens, Greece; Chompoonek Yurayart, Kasetsart University, Bangkok, Thailand; Rosely Maria Zancopé-Oliveira, Instituto de Pesquisa Clínica Evandro Chagas (IPEC), Fundação Oswaldo Cruz, FIOCRUZ, Rio de Janeiro, Brazil.

DATA AVAILABILITY

The primary and secondary fungal barcodes generated and analyzed for this study are available in the ISHAM Barcoding Database (http://its.mycologylab.org/).

AUTHOR CONTRIBUTIONS

WM, SC, and TS conceived the study. WM coordinated the study and supervised the molecular and data analysis. All members of the ISHAM working group for "Barcoding of Medical Fungi" performed the sample collections. MH and LI conducted the sample processing, all molecular studies, and data analysis. MH, LI, and WM performed the data interpretation. All authors wrote and corrected the manuscript.

FUNDING

This study was supported by the National Health and Medical Research Council of Australia (NH & MRC) grant (#APP1121936) to WM, SC, and TS, and Western Sydney Local Health District Research and Education Network Research Grant Scheme to WM, TS, SC, and LI. TS is a Sydney Medical School Foundation Fellow.

ACKNOWLEDGMENTS

The authors thank all members of the ISHAM working group for "Barcoding of Medical Fungi" for the contribution of strains and barcode sequences.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.01647/full#supplementary-material

REFERENCES

Agatha, D., Krishnan, K. U., Dillirani, V. A., and Selvi, R. (2014). Invasive lung infection by *Scedosporium apiospermum* in an immunocompetent individual. *Indian J. Pathol. Microbiol.* 57, 635–637. doi: 10.4103/0377-4929.142716

Allen, P., Koka, R., Kleinberg, M. E., and Baer, M. R. (2013). *Scedosporium apiospermum* soft tissue infection as the initial presentation of acute myeloid leukemia: a case report. *J. Clin. Oncol.* 31, e98–e100. doi: 10.1200/jco.2012.43.9489

Armstrong-James, D., Meintjes, G., and Brown, G. D. (2014). A neglected epidemic: fungal infections in HIV/AIDS. *Trends Microbiol.* 22, 120–127. doi: 10.1016/j.tim.2014.01.001

Balajee, S. A., Sigler, L., and Brandt, M. E. (2007). DNA and the classical way: identification of medically important molds in the 21st century. *Med. Mycol.* 45, 475–490. doi: 10.1080/13693780701449425

Barley, A. J., and Thomson, R. C. (2016). Assessing the performance of DNA barcoding using posterior predictive simulations. *Mol. Ecol.* 25, 1944–1957. doi: 10.1111/mec.13590

Beguin, H., Pyck, N., Hendrickx, M., Planard, C., Stubbe, D., and Detandt, M. (2012). The taxonomic status of *Trichophyton quinckeanum* and T. interdigitale revisited: a multigene phylogenetic approach. *Med. Mycol.* 50, 871–882. doi: 10.3109/13693786.2012.684153

Behera, B., Singh, R., IXess, I., Mathur, P., Hasan, F., and Misra, M. C. (2010). *Candida rugosa*: a possible emerging cause of candidaemia in trauma patients. *Infection* 38, 387–393. doi: 10.1007/s15010-010-0044-x

Benedict, K., Richardson, M., Vallabhaneni, S., Jackson, B. R., and Chiller, T. (2017). Emerging issues, challenges, and changing epidemiology of fungal disease outbreaks. *Lancet Infect. Dis.* 17, e403–e411. doi: 10.1016/S1473-3099(17)30443-7

Benson, D. A., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2014). GenBank. *Nucleic Acids Res.* 42, D32–D37. doi: 10.1093/nar/gkw1070

Berkhout, C. M. (1923). *De Schimmelgeslachten Monilia, Oidium, Oospora en Torula*. Utrecht: Rijksuniversiteit Utrecht.

Bitar, D., Lortholary, O., Le Strat, Y., Nicolau, J., Coignard, B., Tattevin, P., et al. (2014). Population-based analysis of invasive fungal infections, France, 2001–2010. *Emerg. Infect. Dis.* 20, 1149–1155. doi: 10.3201/eid2007.140087

Brown, G. D., Denning, D. W., Gow, N. A., Levitz, S. M., Netea, M. G., and White, T. C. (2012). Hidden killers: human fungal infections. *Sci. Transl. Med.* 4:165rv113. doi: 10.1126/scitranslmed.3004404

Cadena, J., Thompson, G. R. III, and Patterson, T. F. (2016). Invasive Aspergillosis: current strategies for diagnosis and management. *Infect. Dis. Clin. North Am.* 30, 125–142. doi: 10.1016/j.idc.2015.10.015

Ceccarelli, L., Calisti, G., Delle Rose, D., Ricciardi, A., Maffongelli, G., Sordillo, P., et al. (2012). Dapsone hypersensitivity syndrome complicated by scedosporium apiospermum pneumonia in an immunocompetent patient. *Infection* 40, 459–462. doi: 10.1007/s15010-011-0225-2

Chen, S. C., Meyer, W., and Sorrell, T. C. (2014). *Cryptococcus gattii* infections. *Clin. Microbiol. Rev.* 27, 980–1024. doi: 10.1128/CMR.00126-13

Ciardo, D. E., Schar, G., Bottger, E. C., Altwegg, M., and Bosshard, P. P. (2006). Internal transcribed spacer sequencing versus biochemical profiling for identification of medically important yeasts. *J. Clin. Microbiol.* 44, 77–84. doi: 10.1128/JCM.44.1.77-84.2006

Cole, D. C., Govender, N. P., Chakrabarti, A., Sacarlal, J., and Denning, D. W. (2017). Improvement of fungal disease identification and management: combined health systems and public health approaches. *Lancet Infect. Dis.* 17, e412–e419. doi: 10.1016/S1473-3099(17)30308-0

Collins, R. A., Boykin, L. M., Cruickshank, R. H., and Armstrong, K. F. (2012). Barcoding's next top model: an evaluation of nucleotide substitution models for specimen identification. *Methods Ecol. Evol.* 3, 457–465. doi: 10.1111/j.2041-210X.2011.00176.x

Cortez, K. J., Roilides, E., Quiroz-Telles, F., Meletiadis, J., Antachopoulos, C., Knudsen, T., et al. (2008). Infections caused by *Scedosporium* spp. *Clin. Microbiol. Rev.* 21, 157–197. doi: 10.1128/CMR.00039-07

Cruysmans, C., Rodriguez-Villalobos, H., Fomekong, E., Dumitriu, D., Nassogne, M. C., and Van der Linden, D. (2015). Epidural abscess caused by *Scedosporium apiospermum* in an immunocompetent child. *Pediatr. Infect. Dis. J.* 34, 1277–1278. doi: 10.1097/INF.0000000000000864

de Hoog, G. S., Guarro, J., Gené, J., and Figueras, A. A. (2014). *Atlas of Clinical Fungi*. Netherlands: Centraalbureau voor Schimmelcultures (CBS).

Denning, D. W. (2016). Minimizing fungal disease deaths will allow the UNAIDS target of reducing annual AIDS deaths below 500 000 by 2020 to be realized. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 371:0468. doi: 10.1098/rstb.2015.0468

Douglass, A. P., Offei, B., Braun-Galleani, S., Coughlan, A. Y., Martos, R., Ortiz-Merino, R. A., et al. (2018). *Population genomics* shows no distinction between pathogenic *Candida krusei* and environmental *Pichia kudriavzevii*: One species, four names. *PLoS Pathog.* 14:e1007138. doi: 10.1371/journal.ppat.1007138

Dromer, F., Mathoulin-Pelissier, S., Fontanet, A., Ronin, O., Dupont, B., and Lortholary, O. (2004). Epidemiology of HIV-associated cryptococcosis in France (1985-2001): comparison of the pre- and post-HAART eras. *Aids* 18, 555–562. doi: 10.3201/eid2007.140087

Ferrer, C., Colom, F., Frases, S., Mulet, E., Abad, J., and Alio, J. (2001). Detection and identification of fungal pathogens by PCR and by ITS2 and 5.8S ribosomal DNA typing in ocular infections. *J. Clin. Microbiol.* 39, 2873–2879. doi: 10.1128/JCM.39.8.2873-2879.2001

Gardes, M., and Bruns, T. D. (1993). ITS primers with enhanced specificity for basidiomycetes - application to the identification of mycorrhizae and rusts. *Mol. Ecol.* 2, 113–118. doi: 10.1111/j.1365-294X.1993.tb00005.x

Gerrits van den Ende, A. H. G., and de Hoog, G. S. (1999). Va*riability and molecular diagnostics of the neurotropic species *Cladophialophora bantiana*. *Stud. Mycol.* 43, 151–162.

Gilgado, F., Cano, J., Gene, J., Sutton, D. A., and Guarro, J. (2008). Molecular and phenotypic data supporting distinct species statuses for *Scedosporium apiospermum* and *Pseudallescheria boydii* and the proposed new species *Scedosporium dehoogii*. *J. Clin. Microbiol.* 46, 766–771. doi: 10.1128/JCM.01122-07

Guitard, J., Angoulvant, A., Letscher-Bru, V., L'Ollivier, C., Cornet, M., Dalle, F., et al. (2013). Invasive infections due to *Candida norvegensis* and *Candida inconspicua*: report of 12 cases and review of the literature. *Med. Mycol.* 51, 795–799. doi: 10.3109/13693786.2013.807444

Ha, M. V., Choy, M. S., McCoy, D., Fernandez, N., and Suh, J. S. (2018). *Candida catenulata* candidaemia and possible endocarditis in a cirrhotic patient successfully de-escalated to oral fluconazole. *J. Clin. Pharm. Ther.* 43, 910–913. doi: 10.1111/jcpt.12728

Hawksworth, D. L. (2011). A new dawn for the naming of fungi: impacts of decisions made in Melbourne in July 2011 on the future publication and regulation of fungal names. *IMA Fungus* 2, 155–162. doi: 10.5598/imafungus.2011.02.02.06

Hebert, P. D., Cywinska, A., Ball, S. L., and deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 270, 313–321. doi: 10.1098/rspb.2002.2218

Irinyi, L., Lackner, M., de Hoog, G. S., and Meyer, W. (2016). DNA barcoding of fungi causing infections in humans and animals. *Fungal Biol.* 120, 125–136. doi: 10.1016/j.funbio.2015.04.007

Irinyi, L., Serena, C., Garcia-Hermoso, D., Arabatzis, M., Desnos-Ollivier, M., Vu, D., et al. (2015). international society of human and animal mycology (ISHAM)-ITS reference DNA barcoding database-the quality controlled standard tool for routine identification of human and animal pathogenic fungi. *Med. Mycol.* 53, 313–337. doi: 10.1093/mmy/myv008

Johnson, L. S., Shields, R. K., and Clancy, C. J. (2014). Epidemiology, clinical manifestations, and outcomes of Scedosporium infections among solid organ transplant recipients. *Transpl Infect Dis.* 16, 578–587. doi: 10.1111/tid.12244

Khunnamwong, P., Lertwattanasakul, N., Jindamorakot, S., Limtong, S., and Lachance, M. A. (2015). Description of *Diutina* gen. nov., *Diutina siamensis*, f.a. sp. nov., and reassignment of *Candida catenulata*, *Candida mesorugosa*, *Candida neorugosa*, *Candida pseudorugosa*, *Candida ranongensis*, *Candida rugosa* and *Candida scorzettiae* to the genus Diutina. *Int. J. Syst. Evol. Microbiol.* 65, 4701–4709. doi: 10.1099/ijsem.0.000634

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16, 111–120. doi: 10.1007/BF01731581

Kneale, M., Bartholomew, J. S., Davies, E., and Denning, D. W. (2016). Global access to antifungal therapy and its variable cost. *J. Antimicrob. Chemother.* 71, 3599–3606. doi: 10.1093/jac/dkw325

Kõljalg, U., Nilsson, R. H., Abarenkov, K., Tedersoo, L., Taylor, A. F. S., Bahram, M., et al. (2013). Towards a unified paradigm for sequence-based identification of fungi. *Mol. Ecol.* 22, 5271–5277. doi: 10.1111/mec.12481

Kontoyiannis, D. P., Marr, K. A., Park, B. J., Alexander, B. D., Anaissie, E. J., Walsh, T. J., et al. (2010). Prospective surveillance for invasive fungal infections in hematopoietic stem cell transplant recipients, 2001-2006: overview of the transplant-associated infection surveillance network (TRANSNET) database. *Clin. Infect. Dis.* 50, 1091–1100. doi: 10.1086/651263

Kubisiak-Rzepczyk, H., Gil, L., Zawirska, A., Kubisiak-Michalska, A., Mol, A., Reich, A., et al. (2013). *Scedosporium prolificans* fungaemia in a patient with acute lymphoblastic leukaemia. *J. Mycol. Med.* 23, 261–264. doi: 10.1016/j.mycmed.2013.08.003

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Kurtzman, C., and Robnett, C. (1998). Identification and phylogeny of ascomycetous yeasts from analysis of nuclear large subunit (26S) ribosomal DNA partial sequences. *Antonie Van Leeuwenhoek* 73, 331–371. doi: 10.1023/A:1001761008817

Kurtzman, C. P., Fell, J. W., and Boekhout, T. (2011). *The Yeasts: A Taxonomic Study*. Amsterdam: Elsevier.

Kurtzman, C. P., Robnett, C. J., and Basehoar-Powers, E. (2008). Phylogenetic relationships among species of *Pichia*, *Issatchenkia* and *Williopsis* determined from multigene sequence analysis, and the proposal of Barnettozyma gen. nov., *Lindnera gen.* nov. and *Wickerhamomyces gen.* nov. *FEMS Yeast Res.* 8, 939–954. doi: 10.1111/j.1567-1364.2008.00419.x

Lackner, M., Hagen, F., Meis, J. F. Gerrits van den Ende, A. H., Vu, D., Robert, V., et al. (2014). Susceptibility and diversity in the therapy-refractory genus scedosporium. *Antimicrob. Agents Chemother.* 58, 5877–5885. doi: 10.1128/AAC.03211-14

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., et al. (2007). Clustal W and clustal X version 2.0. *Bioinformatics* 23, 2947–2948. doi: 10.1093/bioinformatics/btm404

Librado, P., and Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25, 1451–1452. doi: 10.1093/bioinformatics/btp187

Lu, Q., Gerrits van den Ende, A. H., Bakkers, J. M., Sun, J., Lackner, M., Najafzadeh, M. J., et al. (2011). Identification of *Pseudallescheria* and *Scedosporium* species by three molecular methods. *J. Clin. Microbiol.* 49, 960–967. doi: 10.1128/JCM.01813-10

Maschmeyer, G. (2006). The changing epidemiology of invasive fungal infections: new threats. *Int. J. Antimicrob. Agents* 27(Suppl. 1), 3–6. doi: 10.1016/j.ijantimicag.2006.03.006

Meyer, C. P., and Paulay, G. (2005). DNA barcoding: error rates based on comprehensive sampling. *PLoS Biol.* 3:e422. doi: 10.1371/journal.pbio.0030422

Meyer, W., Irinyi, L., Hoang, M. T. <I. >V. </I. >, Robert, V., Garcia-Hermoso, D., Desnos-Ollivier, M., et al. (2018). Database establishment for the secondary fungal DNA barcode-translational elongation factor 1α (TEF1α). *Genome* 62, 160–169. doi: 10.1139/gen-2018-0083

Minces, L. R., Ho, K. S., Veldkamp, P. J., and Clancy, C. J. (2009). Candida rugosa: a distinctive emerging cause of candidaemia. A case report and review of the literature. *Scand. J. Infect Dis.* 41, 892–897. doi: 10.3109/00365540903161531

Ming, C., Huang, J., Wang, Y., Lv, Q., Zhou, B., Liu, T., et al. (2019). Revision of the medically relevant species of the yeast genus Diutina. *Med. Mycol.* 57, 226–233. doi: 10.1093/mmy/myy001

Morris, A., Lundgren, J. D., Masur, H., Walzer, P. D., Hanson, D. L., Frederick, T., et al. (2004). Current epidemiology of Pneumocystis pneumonia. *Emerg. Infect. Dis.* 10, 1713–1720. doi: 10.3201/eid1010.030985

Nagarathnamma, T., Chunchanur, S. K., Rudramurthy, S. M., Vineetha, K. R., Ramamurthy, K., Joseph, J., et al. (2017). Outbreak of *Pichia kudriavzevii* fungaemia in a neonatal intensive care unit. *J. Med. Microbiol.* 66, 1759–1764. doi: 10.1099/jmm.0.000645

O'Donnell, K. (1993). "*Fusarium* and its near relatives," in *The Fungal Holomorph: Mitotic, Meiotic and Pleomorphic Speciation in Fungal Systematics*, eds D. R. Reynolds and J. W. Taylor (Wallingford: CAB International), 225–233.

Padovan, A. C., Melo, A. S., and Colombo, A. L. (2013). Systematic review and new insights into the molecular characterization of the *Candida rugosa* species complex. *Fungal Genet. Biol.* 61, 33–41. doi: 10.1016/j.fgb.2013.10.007

Pappas, P. G., Alexander, B. D., Andes, D. R., Hadley, S., Kauffman, C. A., Freifeld, A., et al. (2010). Invasive fungal infections among organ transplant recipients: results of the transplant-associated infection surveillance network (TRANSNET). *Clin. Infect. Dis.* 50, 1101–1111. doi: 10.1086/651262

Patterson, T. F. (2005). Advances and challenges in management of invasive mycoses. *Lancet* 366, 1013–1025. doi: 10.1016/S0140-6736(05)67381-3

Pfaller, M. A., Diekema, D. J., Colombo, A. L., Kibbler, C., Ng, K. P., Gibbs, D. L., et al. (2006a). *Candida rugosa*, an emerging fungal pathogen with resistance to azoles: geographic and temporal trends from the ARTEMIS DISK antifungal surveillance program. *J. Clin. Microbiol.* 44, 3578–3582. doi: 10.1128/JCM.00863-06

Pfaller, M. A., Pappas, P. G., and Wingard, J. R. (2006b). Invasive fungal pathogens: current epidemiological trends. *Clin. Infect. Dis.* 43(Suppl. 1), S3–S14. doi: 10.1086/504490

Radosavljevic, M., Koenig, H., Letscher-Bru, V., Waller, J., Maloisel, F., Lioure, B., et al. (1999). *Candida catenulata* fungemia in a cancer patient. *J. Clin. Microbiol.* 37, 475–477.

Rainer, J., and de Hoog, G. S. (2006). Molecular taxonomy and ecology of *Pseudallescheria*, *Petriella* and *Scedosporium prolificans* (Microascaceae) containing opportunistic agents on humans. *Mycol. Res.* 110(Pt 2), 151–160. doi: 10.1016/j.mycres.2005.08.003

Rajasingham, R., Smith, R. M., Park, B. J., Jarvis, J. N., Govender, N. P., Chiller, T. M., et al. (2017). Global burden of disease of HIV-associated cryptococcal meningitis: an updated analysis. *Lancet Infect. Dis.* 17, 873–881. doi: 10.1016/S1473-3099(17)30243-8

Rehner, S. A., and Buckley, E. (2005). A Beauveria phylogeny inferred from nuclear ITS and EF1-alpha sequences: evidence for cryptic diversification and links to *Cordyceps* teleomorphs. *Mycologia* 97, 84–98. doi: 10.3852/mycologia.97.1.84

Sanclemente, G., Marco, F., Cervera, C., Hoyo, I., Colmenero, J., Pitart, C., et al. (2015). *Candida norvegensis* fungemia in a liver transplant recipient. *Rev. Iberoam. Micol.* 32, 115–117. doi: 10.1016/j.riam.2013.11.005

Schelenz, S., Barnes, R. A., Barton, R. C., Cleverley, J. R., Lucas, S. B., and Kibbler, C. C. (2015). British society for medical mycology best practice recommendations for the diagnosis of serious fungal diseases. *Lancet Infect. Dis.* 15, 461–474. doi: 10.1016/S1473-3099(15)70006-X

Schoch, C., Seifert, K., Huhndorf, S., Robert, V., Spouge, J., Levesque, C., et al. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc. Natl. Acad. Sci. U.S.A.* 109, 6241–6246. doi: 10.1073/pnas.1117018109

Schoch, C. L., Robbertse, B., Robert, V., Vu, D., Cardinali, G., Irinyi, L., et al. (2014). Finding needles in haystacks: linking scientific names, reference specimens and molecular data for Fungi. *Database* 2014:bau061. doi: 10.1093/database/bau061

Schuster, M. G., Meibohm, A., Lloyd, L., and Strom, B. (2013). Risk factors and outcomes of Candida krusei bloodstream infection: a matched, case-control study. *J. Infect.* 66, 278–284. doi: 10.1016/j.jinf.2012.11.002

Spampinato, C., and Leonardi, D. (2013). *Candida* infections, causes, targets, and resistance mechanisms: traditional and alternative antifungal agents. *BioMed. Res. Int.* 2013:204237. doi: 10.1155/2013/204237

Srivathsan, A., and Meier, R. (2012). On the inappropriate use of Kimura-2-parameter (K2P) divergences in the DNA-barcoding literature. *Cladistics* 28, 190–194. doi: 10.1111/j.1096-0031.2011.00370.x

Stielow, J. B., Lévesque, C. A., Seifert, K. A., Meyer, W., Irinyi, L., Smits, D., et al. (2015). One fungus, which genes? Development and assessment of universal primers for potential secondary fungal DNA barcodes. *Persoonia* 35, 242–263. doi: 10.3767/003158515X689135

Tamm, M., Malouf, M., and Glanville, A. (2001). Pulmonary Scedosporium infection following lung transplantation. *Transpl. Infect. Dis.* 3, 189–194. doi: 10.1034/j.1399-3062.2001.30402.x

Tavanti, A., Davidson, A. D., Gow, N. A. R., Maiden, M. C. J., and Odds, F. C. (2005). Candida orthopsilosis and *Candida metapsilosis* spp. nov. to replace *Candida parapsilosis* groups II and III. *J. Clin. Microbiol.* 43, 284–292. doi: 10.1128/JCM.43.1.284-292.2005

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680. doi: 10.1093/nar/22.22.4673

Vilgalys, R., and Hester, M. (1990). Rapid genetic identification and mapping of enzymatically amplified ribosomal DNA from several *Cryptococcus* species. *J. Bacteriol.* 172, 4238–4246. doi: 10.1128/jb.172.8.4238-4246.1990

Warnock, D. W. (2007). Trends in the epidemiology of invasive fungal infections. *Nippon Ishinkin Gakkai Zasshi* 48, 1–12. doi: 10.3314/jjmm.48.1

White, T. J., Bruns, T., Lee, S., and Taylor, J. W. (1990). "Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics," in *PCR Protocols: a Guide to Methods and Applications*, eds M. A. Innis, D. H. Gelfandm, J. J. Sninsky, and T. J. White (New York: Academic Press), 315–322. doi: 10.1016/b978-0-12-372180-8.50042-1

Yu, Z., Hu, L., Jiang, M., Ning, H., Xu, C., Li, B., et al. (2013). Dermatic *Scedosporium apiospermum* infection after autologous bone marrow transplantation. *Intern. Med.* 52, 689–693. doi: 10.2169/internalmedicine.52.8532

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Hoang, Irinyi, Chen, Sorrell, the ISHAM Barcoding of Medical Fungi Working Group and Meyer. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Chapter 5
Bioinformatic Improvements in Fungal Metabarcoding

5.1 Linked machine learning classifiers improve species classification of fungi when using error-prone long-reads on extended metabarcodes

The manuscript was submitted to mBio and is undergoing review for resubmission. Hoang M.T.V. was involved in providing fungal material and DNA and manuscript preparation and review. Enjes T., Hu Y., Mashford B., and Schwessinger B. designed the experiment and performed the analysis. Hu Y. performed the data acquisition. Irinyi L., Smith L.M., Linde C.C., and Meyer W. provided fungal material and/or DNA. Stone E. and Rathjen J.P. provided guidance on experimental design and data analysis. Eenjes T., Hu Y., and Schwessinger B. wrote the manuscript. Meyer W., Stone E., Rathjen J.P., and Schwessinger B. provided funding. All authors commented on the manuscript and approved the submission.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Minh Thuy Vi Hoang
30th June 2023

As a supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Professor Wieland Meyer
30th June 2023

Tavish Eenjes[1†], Yiheng Hu[1,8†], Laszlo Irinyi[2,3,4], Minh Thuy Vi Hoang[2,3,4], Leon M. Smith[1,9], Celeste C. Linde[1], Wieland Meyer[2,3,4,5,6], Eric A. Stone[1,7], John P. Rathjen[1], Benjamin Mashford[7*], Benjamin Schwessinger[1*]

[1] Research School of Biology, Australian National University, Acton, ACT, 2601, Australia

[2] Molecular Mycology Research Laboratory, Centre for Infectious Diseases and Microbiology, Faculty of Medicine and Health, Sydney Medical School, Westmead Clinical School, The University of Sydney, Camperdown, NSW, 2006, Australia

[3] Sydney Institute for Infectious Diseases, The University of Sydney, Camperdown, NSW, 2006, Australia

[4] Westmead Institute for Medical Research, Westmead, NSW, 2145, Australia

[5] Westmead Hospital (Research and Education Network), Westmead, NSW, 2145, Australia

[6] Curtin Medical School, Curtin University, Perth Bentley, WA 6102, Australia

[7] Biological Data Science Institute, Australian National University, Acton, ACT, 2601, Australia

[8] Current address: Department of Microbial Interactions, IMIT/ZMBP, University of Tübingen, Tübingen, 72074, Germany

[9] Diversity Arrays Technology, Bruce, ACT, 2617, Australia

[†] These authors contributed equally to this work.

[*]Correspondence should be addressed to Benjamin Mashford (benjamin.mashford@anu.edu.au) and Benjamin Schwessinger (benjamin.schwessinger@anu.edu.au)

5.1.1 Abstract

Background: The increased usage of error-prone long-read sequencing for metabarcoding of fungi has not been matched with adequate public databases and concomitant analysis approaches. We address this gap and present a proof-of-concept study for classifying fungal taxa using linked machine learning classifiers. We demonstrate the capability of linked machine learning classifiers to accurately classify species and strains using real-world and simulated fungal ribosomal DNA datasets, including plant and human pathogens. We benchmark our new approach in comparison to current alignment and k-mer based methods based on synthetic mock communities. We also assess real world applications of species identification in complex unlabelled datasets.

Results: Our machine learning approach assigned individual nanopore long-read amplicon sequences to fungal species with high recall rates and low false positive rates. Importantly, our approach successfully distinguished between closely related species and strains when individual read errors were higher than the genetic distance between individual taxa, which the alignment and k-mer methods could not do. The machine learning approach showed an ability to identify key species with high recall rates, even in complex samples of unknown species composition.

Conclusions: A proof of concept machine learning approach using a tree-descent approach on a decision tree of classifiers can identify known taxa with high accuracy, and precisely detect known target species from complex samples with high recall rates. We propose this approach is suitable for detecting the known knowns of pathogens or invasive species in any environment of mostly unknown composition, including agriculture and wild ecosystems.

5.1.2 Keywords

Machine learning, fungi, long-read sequencing, MinION

5.1.3 Background

DNA sequencing is increasingly becoming an important part of detecting, identifying and classifying fungal species, particularly through DNA barcoding [1]. To date, this process mostly involves the use of short, variable regions of DNA that differ between species and are surrounded by conserved regions suitable for 'universal' primers, enabling PCR amplification over a large variety of fungal taxa [2, 3]. The internal transcribed spacer (ITS) region is used as the primary DNA barcode region for fungal diversity studies [4]. This region contains the two variable components, ITS1 and ITS2, which are a combined 550-600 bp long on average [5]. ITS1 and ITS2 are separated by the conserved 5.8S rRNA gene and flanked by the conserved 18S and 28S rRNA genes (Figure 1a). Although these regions offer a targetable region for identifying fungal species, they have limitations that affect the ability to accurately classify fungi especially at lower taxonomic ranks [5, 6]. The length of the complete ITS1/2 region prevents short-read sequencing platforms to use both in combination for taxonomic classification. Furthermore, the limited selection of 'universal' primers in the region can subject taxonomic studies to primer biases [7].

With the advent and increasing use of long-read sequencing, some of the limitations of short-reads can be bypassed [8]. With long-reads, an extended ITS region can be sequenced encompassing ITS1 and ITS2 in addition to the minor variable regions of the 18S and 28S rRNA subunits using a single 'universal' primer set (Figure 1a) [9-12]. Long-read sequencing technologies from Pacific Biosciences (PacBio) have been extensively benchmarked for fungal metabarcode sequencing [13]. Previous work by Tedersoo *et al.* [14] using the PacBio RSII and Sequel instruments suggest long-read sequencing offers higher resolution for

species identification than short-reads. Specifically, they found that using full-length ITS sequences increased identification precision for fungi by 33% compared to using ITS1 or ITS2 alone. Further, PacBio long-reads better reflected the community composition of a template mock community than traditional Illumina MiSeq short-reads [13, 14]. In contrast, Oxford Nanopore Technologies (ONT) products have not been as extensively benchmarked for fungal species identification using metabarcodes [15]. This is despite the fact that the ONT MinION has the unique advantage of being highly portable for field studies.

Here, we apply ONT long-read sequencing to fungal ribosomal DNA amplicons generated with the NS3 and LR6 primers [16], spanning close to 2900 bp in size (Figure 1a). We refer to this amplicon hereafter as the fungal ribosomal DNA region. Nanopore sequencing introduces a relatively high read error of around 10% at the time of conducting our study [15], but this read error has recently been improved to ~5% with the ONT super high accuracy basecalling models. These high error rates make individual reads less suited for species identification using DNA metabarcodes with currently existing sequence clustering, alignment and k-mer based methods because the genetic distance of the variable regions between closely related species are often lower than the per read error rate [17]. In addition, the entries in most fungal DNA barcode databases, such as the NCBI ITS database and UNITE [18] are relatively short, with a median sequence length of 580 bp and 540 bp, respectively. This limits the analysis capacity of long-reads which encompass both ITS sequences and include minor variable regions in both 18S and 28S rRNA (Figure 1a).

The fungal kingdom is diverse, with an estimated 1.5-5 million species globally, performing important ecosystem functions [19]. For example, mycorrhizal fungi play a key role in regulating the nutrient uptake of almost all land plant species [20] and can improve plant growth under adverse conditions [21], while detritus-degrading saprotrophic fungi play an important role in improving soil fertility [22]. However, fungal pathogens can have severe

adverse effects on human and animal health and agriculture. Fungi can cause large-scale biodiversity losses [23, 24], as demonstrated by the near extinction of amphibian taxa by the globally devastating fungal pathogen *Batrachochytrium dendrobatidis* [25], and the local extinction of several *Myrtaceae* tree species by the rust fungus *Austropuccinia psidii* [26]. Fungal plant pathogens cost billions of dollars in global food production annually [27].

The importance of fungi warrants the development of improved sequence-based detection methods for fungi. In our current study we address the shortcomings of current techniques and assess the applicability of novel machine learning-based sequence analysis methods for long-read metabarcodes from the fungal kingdom. We focus on the detection of known knowns – taxa known to be present in a system – in the context of pathogen infections or host-associated samples. We explored the potential of machine learning classifiers as an alternative method for assigning individual error-prone long-reads to taxa, especially for detecting known or likely pathogen candidates. We demonstrate that our approach works on real data and demonstrate that machine learning has a place in the pathogen detection space as a supplementary tool with specific functions, rather than as a replacement for existing methods.

Within the existing machine learning methods for identifying patterns across various data types [28-30], convolutional neural networks (CNNs) are one method especially suited for identifying the deterministic spatial relationships in a DNA sequence [31-34]. We applied a CNN approach to metabarcoding-based fungal species identification using a uniquely labelled sequencing dataset of a 2900 bp fungal ribosomal DNA region from 44 individually sequenced fungal taxa. This included species and strains associated with plants, from environmental and soil samples, and human pathogens. We used real-world datasets to closely mimic real-world sequencing conditions and demonstrate the potential of machine-learning in the pathogen detection space, and further demonstrate that this method

can be used alongside simulated read datasets. We compared our machine learning approach to three commonly used analysis approaches, including alignment and k-mer based methods, on different in house and publicly available databases. We demonstrated its potential advantage on classifying closely related fungal species with few base pairs difference within the fungal ribosomal DNA region. Furthermore, by including a broad list of fungal taxa across the fungal kingdom and a comprehensive and focused group of fungal taxa related to the target species when training the machine learning models, the approach can identify and detect a known or suspected target species from a complex sample even if not all species present in the sample were included in the training dataset.

5.1.4 Methods

5.1.4.1 Fungal pathogen sample collection, DNA extraction and ITS amplification

We collected different fungal tissue differently for DNA extractions. The tissue collection processes for each fungal taxon are summarized in Supporting Information Table T1.

We used three different DNA extraction methods for all the species in the mock communities. The methods for each species are listed in Supporting Information Table T1. Collectively, we used two commercially available kits: the Qiagen DNeasy Plant Mini Kit (cat. no. 69106) for most plant pathogenic fungi, and the Quick-DNA Fungal/Bacterial Miniprep Kit (cat. no. D6005, Zymo Research) for some human pathogenic fungi following the manufacturer's protocol. We used a phenol chloroform-based DNA extraction method for other human pathogenic and environmental fungi modified from Ferrer et al [35]. Briefly, 100 mg of leaf tissue was homogenized and cells were lysed using cetyl trimethylammonium bromide (CTAB, Sigma-Aldrich) buffer (added RNAse T1, Thermo Fisher, 1,000 units per 1750 µL), followed by a phenol:chloroform:isoamyl alcohol (25:24:1, Sigma-Aldrich) extraction to remove protein and lipids. DNA was precipitated with 700 µL of isopropanol, washed with 1

ml of 70% ethanol, dried for 5 minutes at room temperature, and resuspended in 50 μL TE buffer containing 10 mM Tris and 1 mM EDTA at pH 8. For the human clinical sample and the field infected wheat sample, we directly used the DNA described in the original article [36, 37] for PCR amplification. The quality and average size of genomic DNA were visualized by gel electrophoresis with a 1% agarose gel for 1 hour at 100 volts. DNA was quantified by NanoDrop and Qubit (Life Technologies) according to the manufacturer's protocol.

We used the NS3 (5' GCAAGTCTGGTGCCAGCAGCC '3) and LR6 (5' CGCCAGTTCTGCTTACC '3) primers [16] to generate the fungal ribosomal DNA fragment of all samples, and the EF1-983F (5' GCYCCYGGHCAYCGTGAYTTYAT '3) and EF1-2218R (5' ATGACACCRACRGCRACRGTYTG '3) primers [16] were used to sequence a secondary region, the fungal *elongation factor 1 alpha* region, which was not used for assessing the machine learning method. We used the New England Biolabs Q5 High-Fidelity DNA polymerase (NEB #M0515) for the PCR reaction following the manufacturer's protocol. Around 10 – 30 nanograms of DNA were used in each PCR reaction. After PCR, DNA was purified with one volume of Agencourt AMPure XP beads (cat. No. A63881, Beckman Coulter) according to the manufacturer's protocol and stored at 4°C.

5.1.4.2 Library preparation and DNA sequencing using the MinION

DNA sequencing libraries were prepared using Ligation Sequencing 1D SQK-LSK108 and Native Barcoding Expansion (PCR-free) EXP-NBD103 Kits from ONT, as adapted by Hu and Schwessinger [38] from the manufacturer's instructions with the omission of DNA fragmentation and DNA repair. DNA was cleaned up using a 1x volume of Agencourt AMPure XP beads (cat. No. A63881, Beckman Coulter), incubated at room temperature with

gentle mixing for 5 minutess, washed twice with 200 µL fresh 70% ethanol, the pellet allowed to dry for 2 minutes, and DNA eluted in 51 µL nuclease free water and quantified using NanoDrop® (Thermo Fisher Scientific, USA) and Promega Quantus™ Fluorometer (cat. No. E6150, Promega, USA) following the manufacturer's instructions. All DNA samples showed absorbance ratio A260/A280 > 1.8 and A260/A230 > 2.0 from the NanoDrop®. DNA was end-repaired using NEBNext Ultra II End-Repair/ dA-tailing Module (cat. No. E7546, New England Biolabs (NEB), USA) by adding 7 µL Ultra II End-Prep buffer, 3 µL Ultra II End-Prep enzyme mix. The mixture was incubated at 20°C for 10 minutes and 65°C for 10 minutes. A 1x volume (60 µl) Agencourt AMPure XP clean-up was performed, and the DNA was eluted in 31 µL nuclease free water. Barcoding reaction was performed by adding 2 µL of each native barcode and 20 µL NEB Blunt/TA Master Mix (cat. No. M0367, New England Biolabs (NEB), USA) into 18 µL DNA, mixing gently and incubating at room temperature for 10 minutes. A 1x volume (40 µL) Agencourt AMPure XP clean-up was then performed, and the DNA was eluted in 15 µL nuclease free water. Ligation was then performed by adding 20 µL Barcode Adapter Mix (EXP-NBD103 Native Barcoding Expansion Kit, ONT, UK), 20 µL NEBNext Quick Ligation Reaction Buffer, and Quick T4 DNA Ligase (cat. No. E6056, New England Biolabs (NEB), USA) to the 50 µL pooled equimolar barcoded DNA, mixing gently and incubating at room temperature for 10 minutes. The adapter-ligated DNA was cleaned-up by adding a 0.4x volume (40 µL) of Agencourt AMPure XP beads, incubating for 5 minutes at room temperature and resuspending the pellet twice in 140 µL ABB provided in the SQK-LSK108 kit. The purified-ligated DNA was resuspended by adding 15 µL ELB provided in the SQK-LSK108 (ONT, UK) kit and resuspending the beads. The beads were pelleted again, and the supernatant transferred to a new 0.5 mL DNA LoBind tube (cat. No. 0030122348, Eppendorf, Germany).

In total, four independent sequencing reactions were performed on a MinION flow cell (R9.4, ONT) connected to a MK1B device (ONT) operated by the MinKNOW software (version 2.0.2): 11 species for each flowcell. Each flow cell was primed with 1 mL of priming buffer comprising 480 µL Running Buffer Fuel Mix (RBF, ONT) and 520 µL nuclease free water. 12 µL of amplicon library was added to a loading mix including 35 µL RBF, 25.5 µL Library Loading beads (ONT library loading bead kit EXP-LLB001, batch number EB01.10.0012) and 2.5 µL water with a final volume of 75 µL and then added to the flow cell via the SpotON sample port. The "NC_48Hr_sequencing_FLOMIN106_SQK-LSK108" protocol was executed through MinKNOW after loading the library and running for 48 h. Raw fast5 files were processed using Albacore 2.3.1 software (ONT) for basecalling, barcode de-multiplexing and quality filtering (Phred quality (Q) score of > 7) as per the manufacturer's recommendations.

Raw unfiltered fastq files were uploaded into NCBI Short Reads Archive under BioProject PRJNA725648.

5.1.4.3 Processing and manipulation of fungal pathogen reads

The read datasets obtained contained both the fungal ribosomal DNA sequence and that of *elongation factor 1 alpha* (*TEF1α*). A two-step data filtration process was applied to remove the *TEF1α* reads and outlier reads from the fungal ribosomal DNA read set.

Reads were mapped to an in-house database of fungal ribosomal DNA regions to select reads of a similar general structure to the ITS region. This homology-based filter assumes the structure of the fungal ribosomal DNA region is similar between species due to shared ancestry, repeatedly shown to be true [39]. The in-house database was curated from 28 ITS sequences from the NCBI Nucleotide database, from a range of genera across the fungal

kingdom. This process mapped reads using *minimap2* (v2.17), using the map-ont flag. Reads that failed to map were discarded.

The remaining reads were filtered for read length. The expected read length for the fungal ribosomal DNA region varied by species, averaging 2600-3200 bp. As the mean length and spread of successfully filtered reads differed between samples, a 90% confidence interval cut-off around the mean read length was applied. This interval was sufficient to exclude short or very long reads resulting from incomplete or partial homology filtering or errors in the sequencing or basecalling processes.

5.1.4.4 Augmenting read datasets

Some reads were simulated using a consensus sequence and MinION read error profile to ensure all samples had at least 15,000 reads for use in the design of the machine learning classifiers downstream. NanoSim (v2.0.0) [40] was used for one species, *Geotrichum candidum*, to generate an additional 8,782 simulated cDNA reads. These reads were generated using an identical error profile and length spread to the pre-existing non-simulated reads.

An additional simulated dataset was later generated using NanoSim v3.0.0 for use in testing the ability to train models using simulated data based on references available online, without losing the ability to accurately classify reads from real-world data. A total of 35,000 additional reads were simulated for *Aspergillus flavus* to ensure the maximum number of reads that could be processed using our approach was available. These reads were generated using an error profile identical to the pre-existing non-simulated reads but were based on a reference consensus sequence for *A. flavus*.

5.1.4.5 Generating consensus sequences for each species

The consensus sequence, an aggregate sequence formed from the comparison of multiple sequences that represents the 'true' sequence, was generated using 200 randomly subsampled filtered reads for each sample. Primer sequences were removed using *Mothur* v1.44.11 [41], an alignment file was generated using *muscle* v3.8.1551 [42] and the consensus sequence was generated from this file using *EMBOSS cons* v6.6.0.0 [43]. These consensus sequences were used to confirm taxonomy when species were initially unclear by mapping to the UNITE, SILVA and NCBI databases.

5.1.4.6 Determining the relationships between samples

Using the taxonomic information available for each sample in MycoBank and the results of a BLAST search with the generated consensus sequences, a cladogram was designed to show the relationships between samples based on major taxonomic ranks. A machine learning classifier would be required at each point where two or more samples split on the cladogram (a node) to distinguish between samples for each read, with the cladogram functioning as a decision tree with the node-situated classifiers determining the path of tree descent. Individual strains of the same species were counted as separate species for this cladogram.

5.1.4.7 Creation of neural network classifiers to distinguish between samples

A convolutional neural network (CNN) was chosen as the most appropriate type of machine learning classifier due to its ability to use the spatial relationships between data features in the reads, such as the distance between ITS and other variable groups. CNNs are capable of learning from both minor variation and higher-order features, which is of particular importance given the high read error of nanopore reads. While we did compare a naïve Bayesian and a fully connected multilayer perceptron model to the CNN approach, these

produced quite poor results in comparison, and previous studies using short-read k-mers had shown CNNs to provide an advantage over the existing Bayesian Ribosomal Database Project and other machine learning classifier types [34]. Given this, we selected CNNs to be the basis of our decision tree moving forward.

CNNs work best when there is a balanced number of items in each classification class. As such, for each multiclass node on the cladogram, an equal number of reads were subsampled from each group of samples that would be represented in the node. So, for machine learning classifiers distinguishing between species, each species present contributes an equal number of reads, while at the kingdom level, each phylum contributes an equal number of reads, with said reads being distributed equally amongst all species belonging to that phylum. The number of reads subsampled was based on the largest number of reads available for each sample, with a maximum of 35,000 reads due to computational processing limitations. For each read subsampled, the nucleotide sequence was converted to a numeric sequence, where A, C, G, and T became 0, 1, 2, and 3, respectively. This integer encoding produced the same result as using one-hot encoding to multiple dimensions. As not all sequences were of equal length, but an equal length was required to avoid sequence length being a distinguishing factor in the classifier, all sequences were padded out to a length of 5,000 bp. The padding used a value of 4 to avoid the padding data from affecting the identification of key features for classification.

Each read was assigned a label representing the output class it would belong to in the one-hot format. Labelled reads were then separated into a training set and a test set. The training set contained 85% of the reads, and was used to train the machine learning classifiers, while the test set contained the remaining 15% of labelled reads and was used to test the efficacy of said classifiers on similar data that the classifier had not previously encountered. This was done using the leave out N sequences approach rather than leaving out N taxa, which is more

appropriate to represent the 'dark matter' taxa, as we wanted to focus on detecting the known knowns rather than the potential unknowns [44]. The neural network was created using the Sequential classifier of the *Keras* framework for neural networks [45], containing five layers of neurons and trained across 10 epochs.

Specific details for the design of the machine learning classifiers and the required software packages for machine learning and other analyses can be found at https://github.com/teenjes/fungal_ML.

5.1.4.8 Evaluation of the machine learning classifiers

The test set was used to assess the accuracy of the various machine learning classifiers. As the test set data was labelled, the expected outcome for each read was known, and could be compared to the output of the machine learning classifier. The accuracy, or classification rate, of these classifiers was the proportion of reads in the test set for whom the prediction of the machine learning classifier, as determined by the highest confidence score, matched the expected outcome. This is equivalent to the recall rate [1], where matches to the expected outcome were true positives and matches outside this outcome were false negatives.

$$Recall\ Rate = \frac{True\ Positives}{True\ Positives + False\ Negatives} \qquad [1]$$

5.1.4.9 Chaining machine learning classifiers into a decision tree

When seeking to identify members of a specific taxon in a community, where the members are not immediately obvious from the species' name, it is useful to have samples classified at each taxonomic rank. A singular classifier would require excessive computational power to do this. As such, we chained the machine learning classifiers together into a decision tree with a tree descent approach, based on the cladogram of the species present in our sample. The most confident outcome of the machine learning classifier at one taxonomic rank would

be used to decide the path along the decision tree. This path could either lead into another machine learning classifier, if the path diverged again, or lead all the way down to the species level with the same confidence.

5.1.4.10 Alternative methods for fungal pathogen read classification

We used two different commonly used methods for fungal pathogen metabarcode classification to benchmark the linked machine learning model approach: an alignment-based method in *minimap2*; and a k-mer-based method in *Kraken2*. To compare these methods, we generated an *in silico* mock community from our labelled sequencing data for which we know the ground truth classification for each sequencing read. This mock community contained 13 of the 44 species used to generate the original machine learning decision tree, randomly subsampling 1000 reads from those not previously used for training the machine learning classifiers. Species were selected to focus on species for whom multiple machine learning classifiers would be required, in particular species with populous genera.

For this *minimap2*-based alignment method, two separate databases were used for identification. Firstly, a gold standard database was created in-house to represent the best-case scenario for identification, when all the species present in a sample are also present in the database. This contained the labelled consensus sequences of all 44 species present in the machine learning decision tree, using the consensus sequences already generated from 200 randomly selected filtered reads. The second was a publicly available database of fungal ITS sequences from NCBI (ftp://ftp.ncbi.nlm.nih.gov/refseq/TargetedLoci/Fungi/fungi.ITS.fna.gz, downloaded Feb 2021). *Minimap2* was applied to each of these databases using the map-ont flag. As the alignment tool can return multiple hits if alignment is good enough, only the best hit was taken for each read.

We used *Kraken2* (v2.0.8) to assign the NCBI taxonomic ID for the same 1000 reads of each species as used in the machine learning decision tree. We generated a *Kraken2* NCBI ITS database with the same fasta file downloaded from above. We used the *Kraken2*-build command with the --add-to-library and --build flags. We used the Python pandas module to modify the *Kraken2* output file and the numpy module to calculate the accuracy.

5.1.4.11 Identifying a key species from a complex sample using machine learning

To assess the suitability of machine learning for this problem, we utilised the two complex datasets sampled from fungi-infected sources of unknown compositions: the field infected wheat dataset [36] and the human clinical dataset [37], to create *in silico* mock communities. We randomly subsampled 950 reads from these datasets and spiked in 50 reads from one of two target species with known ground truth: *Aspergillus flavus*, a crop pathogen; and *Candida albicans*, an opportunistic human pathogen present in the human microbiome. This created a total of four 1000-read synthetic communities, two of which paired a target species and dataset from the same source (*A. flavus* with the wheat dataset and *C. albicans* with the clinical dataset) and two communities where the target species would not be expected to be present in the complex dataset unless it had been spiked in. We used the propagated confidence scores for assessing the recall rate for these spiked datasets, where the confidence score at each taxonomic rank was multiplied to give a final overall confidence at the species level.

We created an additional four *in silico* mock communities to assess the change in recall rate and false positive rate [2] as a confidence threshold was applied.

$$False\ Positive\ Rate = \frac{False\ Positives}{False\ Positives + True\ Negatives} \quad [2]$$

Each mock community was created by randomly subsampling 1000 reads from one of *A. flavus* or *C. albicans* samples with known ground truth and adding an additional 1000 randomly subsampled reads from one of the wheat or clinical datasets containing reads of unknown origin. In total, this resulted in four 2000-read *in silico* mock communities. We assumed the datasets with reads of unknown origin did not contain any reads for the target species tested, placing an upper bound on the false positive rate and a lower bound on the true positive rate. Any positive identifications of the target species *A. flavus* or *C. albicans* with a propagated confidence score below the confidence threshold were instead classified as negative identifications.

5.1.4.12 Identifying a key species from a complex sample using classifiers trained on simulated reads

We used the augmented simulated read dataset from *Aspergillus flavus* generated by NanoSim to retrain those five models involved in the decision tree path leading to *A. flavus*, with an identical process to the initial model training. These models were incorporated into the decision tree in place of the original classifiers to test the ability of the decision tree to identify key species from complex samples when trained on simulated data. This used an identical method to the above method for identifying a key species from a complex sample.

5.1.5 Results

5.1.5.1 Design of a decision tree for machine learning classifiers for taxonomic assignment of fungal species

Here we explored the application of machine learning on individual nanopore reads for fungal taxonomic classification. We sequenced the fungal ribosomal DNA region of 44 fungal taxa individually to generate a labelled real-world dataset for which the ground truth is known for each individual read. This makes our dataset uniquely suited for our supervised machine learning approach and for benchmarking studies when comparing this to commonly used classification approaches. Our fungal species dataset included 39 ascomycetous taxa, spanning 19 families and 27 genera, in addition to five basidiomycetes. Given the high sequencing depth of the amplicon sequencing, we performed stringent quality-control steps on all reads to eliminate sequencing biases. We first filtered reads based on homology against a custom-curated database of the fungal ribosomal DNA region, to remove partial reads or reads from other areas of the fungal genome with partial primer binding. We then filtered reads by length to remove reads that were not within a 90% confidence interval around the mean read length of the fungal ribosomal DNA region for each taxon (see Supporting Information Table T1). The *Geotrichum candidum* (former *Galactomyces geotrichum*) sample had too few reads for further processing, hence we complimented those with simulated reads using NanoSim [40]. This resulted in an average of 54,832 ± 35,537 reads available across all taxa. We took a subsample of these quality-controlled reads and split them into a training set and a test set, containing 85% and 15% of the subsampled reads respectively, to be used for training the machine learning classifiers and assessing the performance of the newly generated machine learning classifiers, respectively. We implemented a decision tree to be able to classify individual reads at each taxonomic rank from phylum to species (Figure 1). The taxonomic information for the 44 available

individually sequenced taxa was used to create the cladogram for this decision tree. For simplification, we counted individual strains as separate species for this purpose to test the ability to discern between very closely related taxa. We generated one machine learning classifier for each node in our decision tree, where two or more sub-taxa diverged along the cladogram (Figure 1).

For training each of these classifiers, a balanced dataset was used, such that each possible outcome of the machine learning classifier had an equal number of reads. By giving equal weight to each group within the classifiers, we ensure that the machine learning model can identify the best distinguishing features between groups. These individual classifiers at different taxonomic ranks had a mean recall rate of 97.9 ± 1.1% for correctly classifying reads using the test read dataset. The lowest recall rate belonged to the species-level classifier that distinguished between *Candida* species, with a recall rate of 94.4%. The receiver operator characteristic (ROC) curves demonstrated a good classifier (see the GitHub repository at https://github.com/teenjes/fungal_ML).

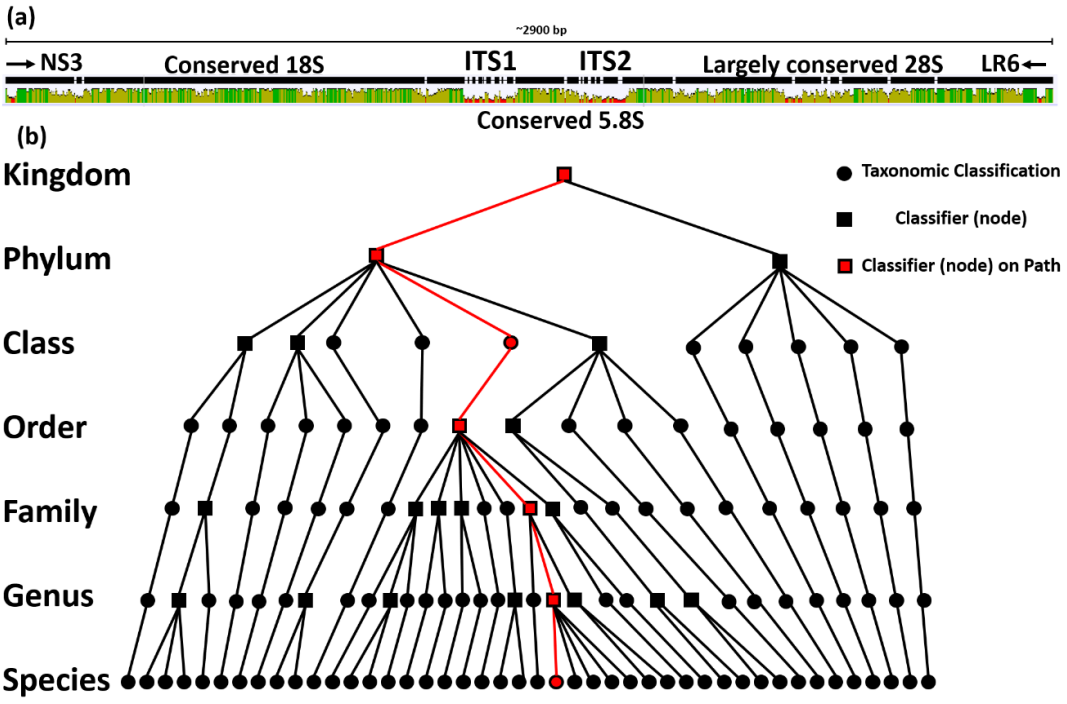


Figure 1: Visualisation of the fungal ribosomal DNA region and machine learning decision tree. (a) The fungal ribosomal DNA region between the NS3 and LR6 primers covers around ~2900 bp. Shown is the alignment of consensus sequences (created with Geneious Prime), with highly conserved regions in green and variable regions in red. **(b)** Taxonomic information for 44 species was used to create a decision tree showing the relationship between samples at each taxonomic rank. Where two or more species shared a common taxon, a machine learning classifier (node) was created to distinguish between those taxa. These classifiers were chained together to create a decision tree, whereby the classification of a read was undertaken by a cascade of classifiers. An example path down the decision tree is shown for *Candida albicans*.

To fully classify a read, we used the cladogram as a decision tree to link individual machine learning classifiers at each taxonomic rank. This allowed us to chain classifiers together to classify a read at each taxonomic rank, moving through the tree from phylum to species assignments. The outcome of a classifier at one taxonomic rank was used to decide the path along the tree, and thus this decision defined which classifier was appropriate for use at the next lower taxonomic rank (Figure 1). We refer to a classifier by the taxonomic rank that it outputs. For example, a species-level classifier takes reads from a specific genus and outputs a species, while a class-level classifier takes reads from a specific phylum and outputs a decision on the taxonomic class of the read. The recall rate of the individual classifiers at different taxonomic ranks can affect the final species-level recall rate for each individual read as it moves through the decision tree. This means that the final species-level recall rate is equal to or worse than the individual species-level classifier's recall rate. Since not every path through the decision tree had a node at each taxonomic rank, the total number of nodes with classifiers is less than the total number of taxa. For example, the basidiomycete species *Puccinia striiformis* f. sp. *tritici* has only two classifiers, at the phylum level and the class level. The latter decides the class classification, but as *P. tritici* is the only species in the class *Pucciniomycetes* in our sequencing dataset, no further nodes were required to distinguish between this species and other taxa present in our dataset. Further, both the *Discula quercina*

and *Tapesia yallundae* species were represented in two samples with different strains, although each was classified as a species for the purposes of this model. In total we trained 22 classifiers to distinguish our 44 fungal taxa.

5.1.5.2 Machine learning methods provide robust classification, especially at the genus and species level

We compared the machine learning decision tree to other, more standard methods for read classification to determine the effectiveness of this technique. We assessed the ability of the other methods at classifying reads across multiple taxonomic ranks because the tiered nature of the decision tree offers the potential to gleam taxonomic information from a read, even when it cannot be confidently classified at the species level. We first looked at the use of a naïve Bayes classifier, however this performed very poorly and was inconsistent in the results produced (data not shown), which is consistent with previous studies [34]. We used two additional classification techniques. We applied *minimap2*, a pairwise alignment-based method designed to be used with long-reads, against a gold-standard custom-curated database generated from the consensus sequences of all 44 species present in the decision tree (gold standard alignment). This is the most appropriate comparison for our machine learning approach because the gold standard and machine learning approaches are directly derived from our sequencing dataset. To compare the machine learning approach with methods where the sequencing data was not used to create the classification database in some way, we applied *minimap2* to a large publicly available database of fungal ITS sequences from NCBI [46, 47] (NCBI alignment), and applied *Kraken2*, a k-mer-based classifier, to the same NCBI database (*Kraken2*).

To compare these methods, an *in silico* mock community was generated from our labelled sequencing data via subsampling for which we know the ground truth classification for each

sequencing read. We selected 13 species from the original 44 taxa, including some that required four or more classifiers to generate the machine learning decision tree, particularly those from populous genera with closely related species. Although all species from this mock community were present in the gold standard database, the NCBI database was missing some genera and species. All of these missing or unclassified taxonomies were recorded as having a recall rate of zero percent, artificially decreasing the quality at lower taxonomic ranks.

Our machine learning decision tree approach maintained a consistently high recall rate across all taxonomic ranks, with a mean species level recall rate of 93.0 ± 2.8% (1σ). Notably, it performed very well for closely related taxa, including the cryptic species *Candida metapsilosis* and *Candida orthopsilosis* and another closely related species *Candida albicans*. The two cryptic *Candida* species (*C. metapsilosis* and *C. orthopsilosis*) had a very high consensus sequence similarity, with a genetic distance of 2.74% (97.26% identity) in our fungal ribosomal DNA target region representing the genetically least distinct species pair. Our machine learning approach achieved species level recall rates of 90.1% and 89.1% for *C. metapsilosis* and *C. parapsilosis*, respectively, even with per read error rates of about 10%. This highlights the strength of our approach.

The gold standard alignment approach also performed very well when compared to the machine learning approach across all taxonomic ranks (Figure 2). The majority of the species were classified with recall rates in excess of 95%. Yet this approach significantly underperformed when trying to differentiate taxa with low genetic distance such as those from the *Candida* genus. As with the machine learning approach, the three *Candida* species were classified with the lowest recall rate at the species level, with *C. albicans, C. metapsilosis* and *C. parapsilosis* being classified with recall rates of 35.8%, 34.0% and 57.5% respectively. These difficulties are also reflected in the overall mean species level recall rate of 76.6 ± 25.5%, which is much lower than our machine learning approach.

Next, we assessed our dataset with alignment and k-mer-based analysis approaches when using the publicly available NCBI database. Overall, NCBI alignment with *minimap2* performed similarly well at higher taxonomic ranks. However, inconsistent or missing naming conventions at the family level and missing or alternate species labels, meant that the overall recall rate was low at the species level, although most of the samples were classified with a high recall rate at the genus level. This low species level recall rate is an artefact created from the choice of database, which is reflected in the similarly poor species level recall rates of the *Kraken2* method. Overall, the k-mer-based *Kraken2* was less accurate than all other methods tested across all taxonomic ranks.

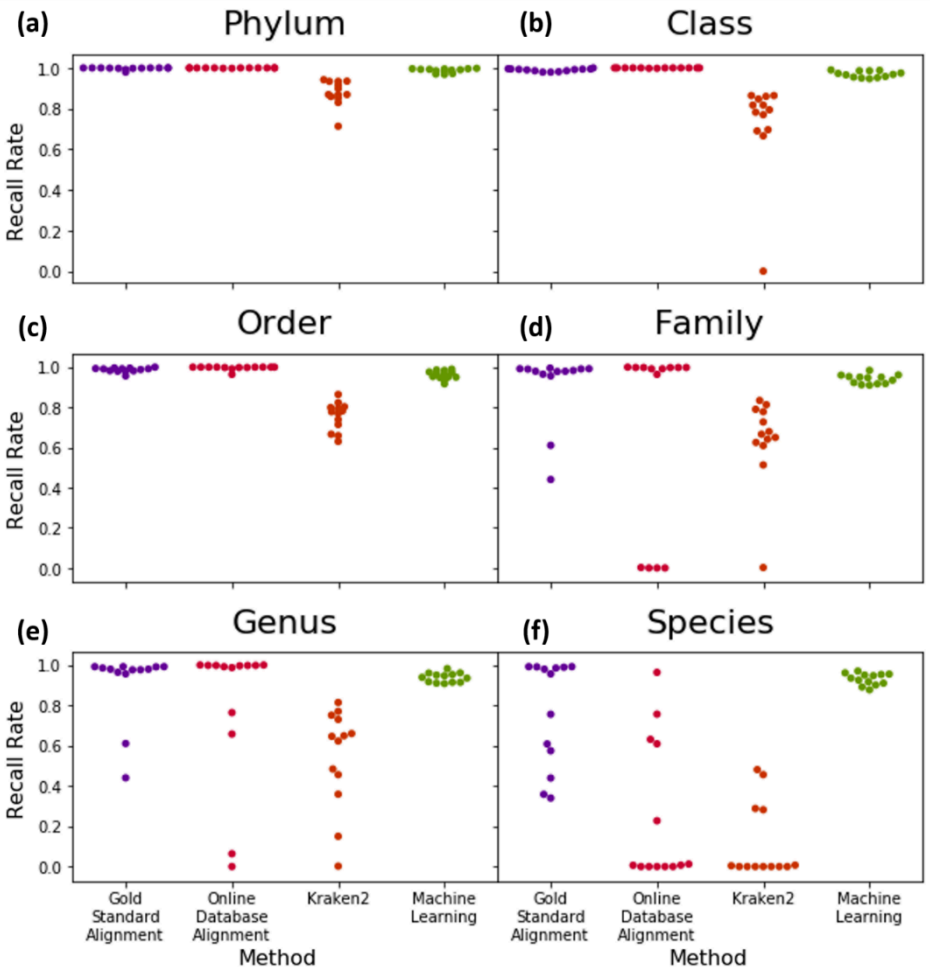


Figure 2: Machine learning-based species identification performs especially well for closely related species. Recall rates of the alignment-based *minimap2* technique, k-mer-based *Kraken2* method and the machine learning decision tree across different

taxonomic ranks **(a-f)**. The *minimap2* technique, as applied to the gold standard database, was successful across most taxonomic ranks, but lower recall rates were recorded for closely related species at the species level **(f)**. Both the *minimap2* and *Kraken2* methods were applied to the NCBI database, and while the *minimap2* NCBI alignment was more accurate across most taxonomic ranks, both showed comparable recall at the species level. The machine learning decision tree approach provided the greatest classification power for closely related species, despite lower recall rates for some distantly related species than the gold standard alignment method.

5.1.5.3 Identifying target species from a complex sample of unknown composition using the machine learning decision tree

A key feature of a species classification tool is its ability to identify a known target species from a complex sample of unknown composition. This is especially important when attempting to identify the presence of a target species, such as a specific pathogen, from a patient or environmental sample.

We generated two additional sequencing datasets of unknown composition from two distinct environments to test the capability of our machine learning decision tree to identify a given target species. These datasets were generated with the same PCR and sequencing protocols as for the individual 44 training species focusing on the fungal ribosomal DNA region. The first dataset was derived from amplicon sequencing of fungi-infected wheat leaves (wheat dataset) and the second was derived from bronchoalveolar wash in a clinical setting (clinical dataset). To each of these sequencing datasets of unknown composition, we spiked *in silico* a known number of reads with known labels as test cases. We choose *Aspergillus flavus*, a crop pathogen, and *Candida albicans*, a human pathogen, to compare two species without and with very close sister species in our linked machine learning classifiers. We then tested recall and false positive rate of our machine learning classifiers using our *in silico* spiked reads, assuming that the original datasets of unknown composition did not contain any reads of either species.

We first plotted the propagated confidence score of the species level classification for all reads in each *in silico* spiked dataset to better understand the behaviour of our machine learning decision tree on samples containing reads of unknown origin (Supporting Information Figure S1). This clearly shows that the propagated confidence scores for reads of unknown origin are far lower than reads of species the classifiers were trained on. We then assessed the recall and false positive rate of the *in silico* spiked datasets at different confidence scores thresholds assuming that the original sample did not include any reads of the *in silico* spiked species (Figure 3). Increasing the thresholds reduced the recall and false positive rate in both cases. For *A. flavus*, the recall rate remained above 90% until the confidence threshold reached 0.9, and the false positive rate was consistently low across both the clinical and wheat datasets with reads of unknown origin. A confidence threshold of 0.85 resulted in a high recall rate of 0.917, while maintaining a low false positive rate of just one percent. For *C. albicans*, not using a confidence threshold at all resulted in a recall rate of 87.7% and false positive rate of 11.7%. However, by using a confidence threshold of 0.85, the recall rate was only decreased to 72.4% while reducing the false positive rate to only 1.7% in the clinical dataset. We recommend this confidence score threshold of 0.85 as suitable for retaining a high recall rate while achieving a low false positive rate, even for a member of a difficult-to-distinguish genus like *Candida*.

5.1.5.4 Training a machine learning classifier with simulated reads resulted in similar accuracy to those trained on real data

Previously, models were trained using sequence data from pure fungal culture. However, this is not always accessible, and limits the ability for the approach to be extended. This allows for samples with low read counts, such as the *G. candidum* sample, to be augmented to have sufficient read counts for training and enables the rapid extension of the suggested approach

to more species. We simulated 35,000 reads for *A. flavus* based on our reference ribosomal DNA sequence using NanoSim v3.0.0 to test whether training machine learning models using simulated reads would produce the same level of accuracy as training with reads generated by sequencing amplicons generated from fungal reference cultures.

We retrained all models in the decision tree path of *A. flavus* using these simulated reads. None of the resultant models were statistically different from those models trained using real sequence data in terms of precision or recall. Further, they showed similar ROC curves (see the GitHub repository at https://github.com/teenjes/fungal_ML). This suggests machine learning classifiers can be trained using simulated reads and still accurately distinguish between taxa across all taxonomic ranks.

We also used these modes trained using simulated data to test whether the use of simulated training data affected the ability of the system to detect target species from a complex sample. We found no difference in the recall or false positive rates for detecting the target species using these simulated-trained models, with a near identical distribution to that shown in Supporting Information Figure S1, and a recall rate of 0.910 when using a confidence threshold of 0.85 compared to the recall rate of 0.917 when trained on real data.

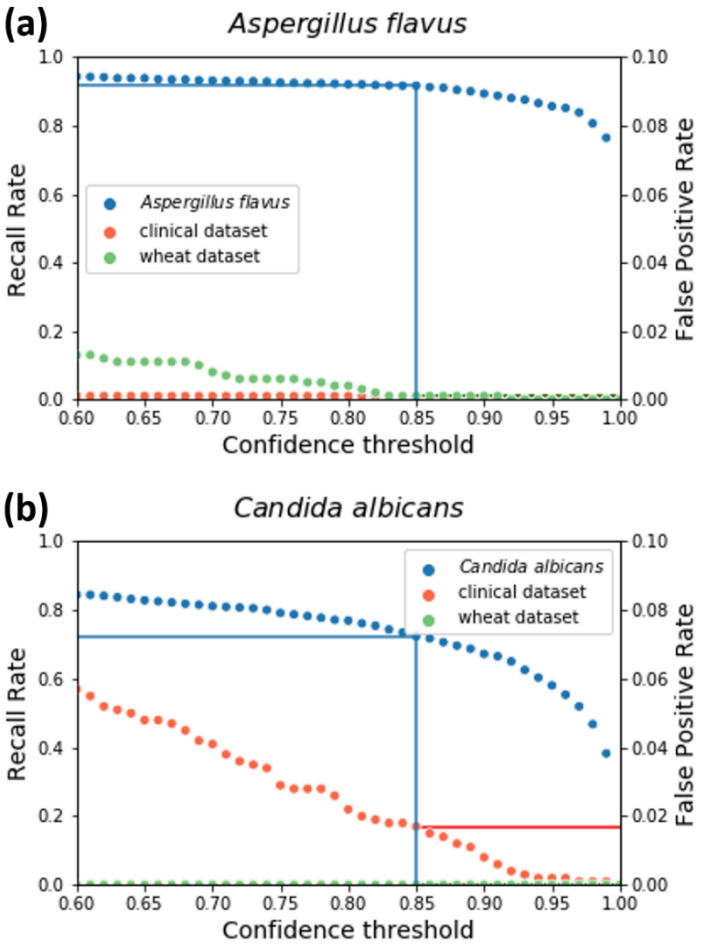


Figure 3: Varying the confidence score threshold affects the recall rate and false positive rate when identifying target species from complex samples of unknown composition. The plots show the recall rate (left axis, blue dots) and false positive rate (right axis, orange or green dots) at varying propagated confidence score thresholds for *Aspergillus flavus* **(a)** and *Candida albicans* **(b)** when spiked into clinical (orange) or wheat (green) datasets. Both plots are based on 2000-read *in silico* spiked samples containing 1000 reads with known labels (*A. flavus* or *C. albicans*) and 1000 reads of unknown origin. For *A. flavus*, a confidence threshold of 0.85 maintains a recall rate of 91.7%, while reducing the false positive rate to 1% for both datasets. For *C. albicans*, the same confidence threshold of 0.85 has a recall rate of 72.4% and reduces the false positive rate below 2% for both datasets.

5.1.6 Discussion

Nanopore sequencing offers portable sequencing platforms and generates long-reads that can cover extended metabarcodes that include more sequence information than the classic metabarcodes from Illumina short-read sequencing [48]. The use of long-read nanopore

sequencing is increasing, and analysis pipelines have been developed for real-time species identification on the MinION, such as the 'What's in my pot' pipeline for complex sample analysis [49]. Yet currently, metabarcode datasets in publicly available databases are limited in barcode length and often do not cover these extended regions, limited only to ITS1 and ITS2 regions of ~580 bps. This poses challenges when classifying error-prone long-reads into different species when the genetic differences within those metabarcode regions are less than the differences from sequencing errors [50].

Here, we implement a novel machine learning approach for species level classification. We used a tree descent approach, with a similar strategy to the IDTAXA taxonomic classifier [51], although the objectives differ. IDTAXA also uses a tree-like approach to classify a subsample of the k-mers into a given taxonomic rank using machine learning, and the user can specify certain confidence thresholds to avoid over-classification when confidence is low [51]. While IDTAXA's goal is to classify unknown reads to a confident taxonomic rank, our approach is more focused on the detection of one or a few target species from mixed samples. As such, we used full-length reads rather than k-mers from the short ITS regions, as used by Vu *et al.* [34], for training the machine learning models. Training these models with long reads is computationally less efficient than with k-mers but is more suitable for detection-focused objectives. The use of only the ITS regions themselves, rather than the extended read used here, may be feasible for detection-focused tasks, but may lose resolution gained from small variations in the otherwise conserved RNA-coding regions that surround the ITS. As such, very closely related taxa may be more difficult to distinguish if only the non-coding ITS is used [34].

Our approach showed great potential in improving the accuracy of sequencing-based fungal classification at a species level. For example, some closely related species, such as those from the *Candida* genus, were readily misclassified with a recall rate lower than 50% using the

gold standard alignment strategy. This is indicative of the problems of alignment-based classification methods for fungi, especially given the relatively high error rate of the nanopore long-reads [52]. Using our machine learning method, the same group of species was classified with recall rates of at least 90% (Figure 2f). This is remarkable given the per-read error rate of 10% for the nanopore reads used, which is much larger than the genetic distance of 2.74% observed between some closely related taxa. With recent improvements in the sequencing and basecalling of nanopore reads reducing the per-read error rate [53], the resolution of our method is likely to increase.

To ensure accurate classification of reads from closely related taxa using our machine learning classifiers, the training of species-level classifiers must include metabarcodes from other taxa closely related to the target for detection. If closely related taxa are not included in training, the likelihood of false positive hits increases as closely related taxa may be identified as the target taxon even when the target taxon is not present. For example, *Candida metapsilosis* may be falsely identified as *Candida albicans* if the model was not trained on both species. As such, accurate species-level identification requires that machine learning classifiers are trained on long metabarcodes from as many fungal species in the same genus as possible. This is especially important when a genus contains both pathogenic and non-pathogenic species. In this way, our approach might be particularly applicable to targeted diagnostic tasks in specific settings, such as detecting fungal pathogens in agriculture [54], forest management [55] and medicine [56], or screening imports for invasive pathogens in aid of border biosecurity [57, 58].

The initial comparisons were based on idealised databases derived from our sequencing dataset, for which metabarcode length and online database entry length were equivalent. Hence, is it expected that analyses based on these idealised databases will outperform other approaches relying on public databases with short reference sequences of uneven length, as

shown by our results (Figure 2). Interestingly, the *Kraken2* approach underperformed compared to the alignment-based approach in our current study. This is consistent with previous studies with nanopore data, where the classification accuracy of *Kraken2* never exceeded that of BLAST, another alignment-based classification program, when using the default 35 bp k-mers [59, 60]. One approach to improve the k-mer-based classification accuracy for error-prone reads would be to use a smaller k-mer length. Another common issue of public fungal databases was that many species were not included in the database or were presented with different taxonomic labels, which resulted in 0% recall rates for some families and species. Changing fungal nomenclature over time can be an issue when using these online databases for classifying fungal metabarcode reads, as the nomenclature is not always updated, leading to outdated or uncorrected taxonomic information persisting in databases [61, 62].

We also tested if our machine learning approach could accurately identify target species in complex samples of unknown composition without having classifiers for all fungal species present in the sample. These complex samples had previously been explored through shogun sequencing [36, 63]. We were able to show that by only training a limited set of classifiers we can detect target species with relatively low false positive and high recall rates in *in silico* spiked datasets with known ground truth of the spiked reads only. By adjusting the confidence score one can decide how many false positives and false negatives one is willing to tolerate. We found a threshold of 0.85 on the propagated confidence score at the species level classification was sufficient to reduce the false positive rate while maintaining high recall rates. Here, the false positive rate was misleadingly high for *Candida* species in the clinical sample and for *Aspergillus* species in the wheat sample. We naively assumed these species were not present in our complex sample save for those reads that had been spiked in. *Candida* species, including *C. albicans*, are common in humans, and similarly *A. flavus* is a

common fungus present on wheat. However, it also demonstrates the need to curate closely related species for training the models, as other species from the *Candida* or *Aspergillus* genera may have been misidentified as the target species from these complex samples.

Further, we found that training the machine learning classifiers using simulated reads, rather than sequenced reads, did not disadvantage the model's ability to distinguish between taxa at any level. It is useful for extending this approach to a larger training set with an increased number of species, particularly for including additional taxa closely related to the target species. We can simulate reads based on publicly available reference sequences to train the classifiers, making this approach accessible even where reference cultures are not readily available.

The taxa used to train the classifiers are flexible and can be changed to suit the user's need including simulation based on publicly deposited genomes of unavailable species. For example, additional species from a specific taxon could be added to reduce false positive detections within this taxa. While many existing classifiers are designed for species classification, and can falsely identify previously unknown species as existing species by over-labelling [44], our approach focused on detection of known target fungi rather than identifying all species present. While unknown taxa absent in the training dataset will be assigned a label, the use of a confidence score threshold can aid in filtering these misclassified unknowns out. The principles behind the application of machine learning to the fungal ribosomal DNA region can be expanded to other barcoding regions, such as *cytochrome c oxidase I* [64] or *elongation factor 1 alpha* [56]. Recent developments by Oxford Nanopore Technologies to improve barcoding accuracy, cost-effectiveness and scalability, such as the recently released Q20+ chemistry, offers promise for expanding our proof-of-concept study to more species using multiple barcodes to improve the species-level resolution and overall classification accuracy [65].

5.1.7 Conclusions

Long-read sequencing can be more accurate for species identification and detection than short reads, offering an advantage for future use [13, 14]. Here, we propose a tangible solution for fungal classification and detection by applying a novel neural network-based machine learning approach on extended fungal ribosomal DNA barcodes using nanopore long-read sequencing. Our machine learning approach can identify target taxa with high accuracy from complex samples of unknown origin, making it applicable to pathogen identification in agriculture, biosecurity, and clinical settings. Our approach performs especially well on closely related species, outcompeting current alignment-based or k-mer-based classification methods for long reads. Future optimisations for real life applications are required in terms of computational efficiency and size of reference training sets.

5.1.8 Declarations

Availability of Data and Materials

The code generated and used for machine learning during the current study is available in the GitHub repository, available at https://github.com/teenjes/fungal_ML. The single-species datasets generated and/or analysed during the current study are available in SRA under BioProject PRJNA725648, available at http://www.ncbi.nlm.nih.gov/bioproject/725648. The complex mixed species ‘clinical’ and ‘wheat’ datasets analysed during the current study are available in SRA.

Funding

This work was supported by computational resources provided by the Australian Government through the National Computational Infrastructure (NCI) under the ANU Merit Allocation

Scheme. YH, ES, JR and BS are supported by The Hermon Slade Foundation grant HSF_17_04, and an Australian NHMRC grant (#APP1121936) to WM.

Author Contributions

TGE, YH, BM, and BS designed the experiments and performed the analysis. YH extracted fungal DNA and performed all sequencing reactions. LL, MTV, LMS, CCL, and WM provided fungal material and/or DNA. ES and JR provided feedback on experimental design and data analysis. WM, ES, JR, and BS provided funding for the project. TGE, YH and BS wrote the manuscript. All authors commented on the manuscript and approved submission.

Acknowledgements

We thank Jen Taylor, and Peter Solomon for suggestions on improving the machine learning models and determining statistics for identifying a species from a complex sample. We thank Peter Solomon and David Jones for providing fungal samples. We thank Andrew Milgate for providing infected wheat leaf material. This work was supported by computational resources provided by the Australian Government through the National Computational Infrastructure (NCI) under the ANU Merit Allocation Scheme.

5.1.9 References

1. Kennedy, P.G., L.C. Cline, and Z. Song, *Probing promise versus performance in longer read fungal metabarcoding.* New Phytol., 2018. **217**(3): p. 973-976.
2. Seifert, K.A., *Progress towards DNA barcoding of fungi.* Mol. Ecol. Resour., 2009. **9**(s1): p. 83-89.
3. Valentini, A., F. Pompanon, and P. Taberlet, *DNA barcoding for ecologists.* Trends Ecol. Evol., 2009. **24**(2): p. 110-117.

4. Schoch, C.L., *et al.*, *Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi.* Proc. Natl. Acad. Sci. U. S. A., 2012. **109**(16): p. 6241-6246.

5. Xu, J., *Fungal DNA barcoding.* Genome, 2016. **59**(11): p. 913-932.

6. Badotti, F., *et al.*, *Effectiveness of ITS and sub-regions as DNA barcode markers for the identification of Basidiomycota (Fungi).* BMC Microbiol., 2017. **17**(1): p. 42.

7. James, T.Y., *et al.*, *Toward a Fully Resolved Fungal Tree of Life.* Annu. Rev. Microbiol., 2020. **74**(1): p. 291-313.

8. Mafune, K.K., *et al.*, *A rapid approach to profiling diverse fungal communities using the MinION™ nanopore sequencer.* Biotechniques, 2020. **68**(2): p. 72-78.

9. Tedersoo, L. and S. Anslan, *Towards PacBio-based pan-eukaryote metabarcoding using full-length ITS sequences.* Environ. Microbiol. Rep., 2019. **11**(5): p. 659-668.

10. Vilgalys, R. and M. Hester, *Rapid genetic identification and mapping of enzymatically amplified ribosomal DNA from several Cryptococcus species.* J. Bacteriol., 1990. **172**(8): p. 4238-46.

11. White, T., J., *et al.*, *Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics*, in *PCR Protocols: A Guide to Methods and Applications*. 1990, Academic Press: New York. p. 315-322.

12. Gardes, M. and T.D. Bruns, *ITS primers with enhanced specificity for basidiomycetes - application to the identification of mycorrhizae and rusts.* Mol. Ecol., 1993. **2**(2): p. 113-118.

13. Castaño, C., *et al.*, *Optimized metabarcoding with Pacific biosciences enables semi-quantitative analysis of fungal communities.* New Phytol., 2020. **228**(3).

14. Tedersoo, L., A. Tooming-Klunderud, and S. Anslan, *PacBio metabarcoding of Fungi and other eukaryotes: errors, biases and perspectives.* New Phytol., 2018. **217**(3): p. 1370-1385.

15. Loit, K., *et al.*, *Relative performance of MinION (Oxford Nanopore Technologies) versus Sequel (Pacific Biosciences) third-generation sequencing instruments in identification of agricultural and forest fungal pathogens.* 2019. **85**(21): p. e01368-19.

16. Raja, H.A., *et al.*, *Fungal identification using molecular tools: A primer for the Natural Products Research Community.* J. Nat. Prod., 2017. **80**(3): p. 756-770.

17. Nilsson, R.H., *et al.*, *Intraspecific ITS variability in the kingdom fungi as expressed in the international sequence databases and its implications for molecular species identification.* Evol. Bioinform. Online, 2008. **4**: p. 193-201.

18. UNITE Community, *UNITE QIIME release for fungi. Version 18.11.2018*. 2019, UNITE Community.

19. Choi, J. and S.-H. Kim, *A genome tree of life for the fungi kingdom.* Proc. Natl. Acad. Sci. U. S. A., 2017. **114**(35): p. 9391-9396.

20. van der Heijden, M.G.A., *et al.*, *Mycorrhizal ecology and evolution: the past, the present, and the future.* New Phytol., 2015. **205**(4): p. 1406-1423.

21. Chen, M., *et al.*, *Beneficial services of arbuscular mycorrhizal fungi – From ecology to application.* Front. Plant Sci., 2018. **9**(1270): p. 1-14.

22. Ning, Q., *et al.*, *Saprotrophic fungal communities in arable soils are strongly associated with soil fertility and stoichiometry.* Appl. Soil Ecol., 2021. **159**: p. 103843.

23. Almeida, F., M.L. Rodrigues, and C. Coelho, *The still underestimated problem of fungal diseases worldwide.* Front. Microbiol., 2019. **10**: p. 214-214.

24. Powell, D., *et al.*, *Genome sequence of the fungus Nannizziopsis barbatae, an emerging reptile pathogen.* Microbiol. Resour. Announc., 2021. **10**(1): p. e01213-20.

25. Lips, K.R., *Overview of chytrid emergence and impacts on amphibians.* Philos. Trans. R. Soc. Lond. B. Biol. Sci., 2016. **371**(1709): p. 20150465.

26. Fensham, R.J., *et al.*, *Imminent extinction of Australian Myrtaceae by fungal disease.* Trends Ecol. Evol., 2020. **35**(7): p. 554-557.

27. Birren, B., *et al.*, *Fungal genome initiative: white paper developed by the fungal research community.* 2002.

28. Tarca, A.L., *et al.*, *Machine learning and its applications to biology.* PLoS Comput. Biol., 2007. **3**(6): p. e116-e116.

29. Libbrecht, M.W. and W.S. Noble, *Machine learning applications in genetics and genomics.* Nat. Rev. Genet., 2015. **16**(6): p. 321-332.

30. Boža, V., B. Brejová, and T. Vinař, *DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads.* PLOS One, 2017. **12**(6): p. e0178751.

31. Flagel, L., Y. Brandvain, and D.R. Schrider, *The unreasonable effectiveness of convolutional neural networks in population genetic inference.* Mol. Biol. Evol., 2019. **36**(2): p. 220-238.

32. Arratia, A. and E. Sepúlveda, *Convolutional neural networks, image recognition and financial time series forecasting, in mining data for financial applications 2019*. 2020, Springer International Publishing: Cham. p. 60-69.

33. Ji, S., *et al.*, *3D Convolutional neural networks for human action recognition.* IEEE Trans. Pattern Anal. Mach. Intell., 2013. **35**(1): p. 221-231.

34. Vu, D., M. Groenewald, and G. Verkley, *Convolutional neural networks improve fungal classification.* Sci. Rep., 2020. **10**(1): p. 12628-12628.

35. Ferrer, C., *et al.*, *Detection and identification of fungal pathogens by PCR and by ITS2 and 5.8S ribosomal DNA typing in ocular infections.* Journal of clinical microbiology, 2001. **39**(8): p. 2873-2879.

36. Hu, Y., *et al.*, *Pathogen detection and microbiome analysis of infected wheat using a portable DNA sequencer.* Phytobiomes J., 2019. **3**(2): p. 92-101.

37. Irinyi, L., *et al.*, *Long-read sequencing based clinical metagenomics for the detection and confirmation of Pneumocystis jirovecii directly from clinical specimens: A paradigm shift in mycological diagnostics.* Medical Mycology, 2019. **58**(5): p. 650-660.

38. Hu, Y. and B. Schwessinger. *Amplicon sequencing using MinION optimized from 1D native barcoding genomic DNA*. 2018 [cited 1-17.

39. Narutaki, S., *et al.*, *Identification of fungi based on the nucleotide sequence homology of their internal transcribed spacer 1 (ITS1) region.* PDA J. Pharm. Sci. Technol., 2002. **56**(2): p. 90-98.

40. Yang, C., *et al.*, *NanoSim: nanopore sequence read simulator based on statistical characterization.* Gigascience, 2017. **6**(4): p. 1-6.

41. Schloss, P.D., *et al.*, *Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities.* Appl. Environ. Microbiol., 2009. **75**(23): p. 7537-7541.

42. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res., 2004. **32**(5): p. 1792-1797.

43. Madeira, F., *et al.*, *The EMBL-EBI search and sequence analysis tools APIs in 2019.* Nucleic Acids Res., 2019. **47**(W1): p. 636-641.

44. Edgar, R.C., *Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences.* PeerJ, 2018. **6**: p. 1-29.

45. Chollet, F., *Keras*. 2015, GitHub.

46. Schoch, C.L., *et al.*, *Finding needles in haystacks: linking scientific names, reference specimens and molecular data for fungi.* Database : the journal of biological databases and curation, 2014. **2014**: p. bau061.

47. Robbertse, B., *et al.*, *Improving taxonomic accuracy for fungi in public sequence databases: applying 'one name one species' in well-defined genera with Trichoderma/Hypocrea as a test case.* Database (Oxford), 2017. **2017**: p. bax072.

48. D'Andreano, S., A. Cuscó, and O. Francino, *Rapid and real-time identification of fungi up to species level with long amplicon nanopore sequencing from clinical samples.* Biol. Methods Protoc., 2020. **6**(1): p. 1-6.

49. Juul, S., *et al.*, *What's in my pot? Real-time species identification on the MinION™.* bioRxiv, 2015: p. 030742.

50. Heeger, F., *et al.*, *Long-read DNA metabarcoding of ribosomal RNA in the analysis of fungi from aquatic environments.* Mol. Ecol. Resour., 2018. **18**(6): p. 1500-1514.

51. Murali, A., A. Bhargava, and E.S. Wright, *IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences.* Microbiome, 2018. **6**(1): p. 140.

52. Hofstetter, V., *et al.*, *The unbearable lightness of sequenced-based identification.* Fungal Divers., 2019. **96**(1): p. 243-284.

53. Marcolungo, L., *et al.*, *Real-time on-site diagnosis of quarantine pathogens in plant tissues by nanopore-based sequencing.* Pathogens, 2022. **11**(2): p. 199.

54. Zou, J., *et al.*, *A primer on deep learning in genomics.* Nat. Genet., 2019. **51**(1): p. 12-18.

55. Luchi, N., R. Ioos, and A. Santini, *Fast and reliable molecular methods to detect fungal pathogens in woody plants.* Appl. Microbiol. Biotechnol., 2020. **104**(6): p. 2453-2468.

56. Ching, T., *et al.*, *Opportunities and obstacles for deep learning in biology and medicine.* J. R. Soc. Interface, 2018. **15**(141): p. 1-47.

57. Bulman, S.R., *et al.*, *Opportunities and limitations for DNA metabarcoding in Australasian plant-pathogen biosecurity.* Australas. Plant Pathol., 2018. **47**(5): p. 467-474.

58. Hu, Y., *et al.*, *Blurred lines: integrating emerging technologies to advance plant biosecurity.* Curr. Opin. Plant Biol., 2020. **56**: p. 127-134.

59. Pearman, W.S., N.E. Freed, and O.K. Silander, *Testing the advantages and disadvantages of short- and long- read eukaryotic metagenomics using simulated reads.* BMC Bioinformatics, 2020. **21**(1): p. 220-220.

60. Hu, Y., *et al.*, *Inferring species compositions of complex fungal communities from long- and short-read sequence data.* mBio, 2022. **13**(2): p. 1-16.

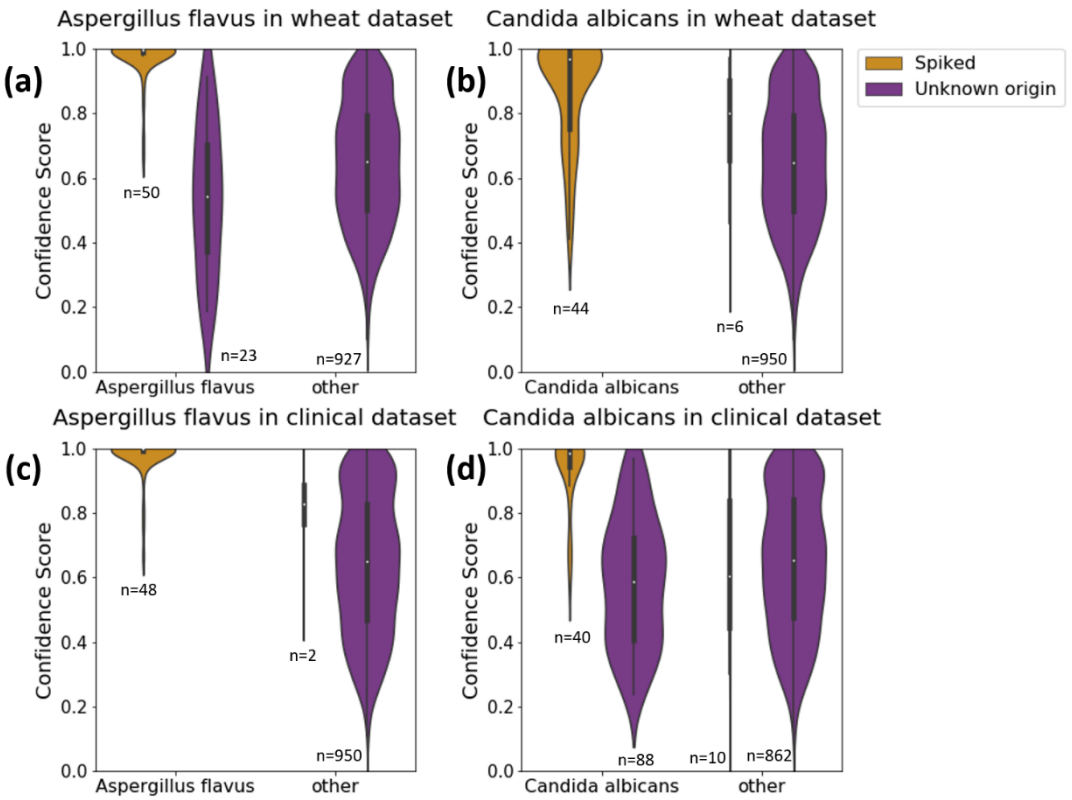
61. Sanford, R.A., *et al.*, *Microbial taxonomy run amok.* Trends Microbiol., 2021. **29**(5): p. 394-404.

62. Lücking, R., *et al.*, *Fungal taxonomy and sequence-based nomenclature.* Nat. Microbiol., 2021. **6**(5): p. 540-548.

63. Irinyi, L., *et al.*, *Long-read sequencing based clinical metagenomics for the detection and confirmation of Pneumocystis jirovecii directly from clinical specimens: A paradigm shift in mycological diagnostics.* Med. Mycol., 2020. **58**(5): p. 650-660.

64. Rodrigues, M.S., K.A. Morelli, and A.M. Jansen, *Cytochrome c oxidase subunit 1 gene as a DNA barcode for discriminating Trypanosoma cruzi DTUs and closely related species.* Parasit. Vectors, 2017. **10**(1): p. 488.

65. Srivathsan, A., *et al.*, *MinION barcodes: biodiversity discovery and identification by everyone, for everyone.* BMC Biol., 2021. **19**(1): p. 217.



Supporting Information Figure S1: The propagated confidence score is markedly different between spiked reads with known labels and reads of unknown origin. The plots show the confidence score of spiked labelled reads (yellow) and reads of unknown origin (purple) at the species level when analysed with our machine learning approach. Labelled reads of species for which we trained classifiers had a markedly better confidence score at the species level, especially when classified into the correct species. Reads of unknown origin had relatively low confidence scores independent of their species level classification. Our 1000-read in silico spike in samples are comprised of 950 reads from a complex sample with reads of unknown origin and 50 in silico spiked reads in samples with known labels from Aspergillus flavus (**a** and **c**) or Candida albicans (**b** and **d**). The complex samples of unknown composition are derived from infected wheat leaves (wheat dataset, **a** and **b**) or from a bronchoalveolar wash taken in a clinical setting (clinical dataset, **c** and **d**). The number of spiked or unspiked reads classified as either the spiked species or any other species is shown.

Supporting Information Table 1: Sample labels, collection methods, DNA extraction methods and read counts before and after two-step data filtering. a) Sabourand dextrose agar (SDA); b) Potato dextrose agar (PDA); c) Yeast extract peptone dextrose (YPD).

Genus label	Species label	Strain/Isolate	Sample Collection	DNA Extraction Method	# Raw Reads	# Homology Filtered	# Length Filtered
Aspergillus	*flavus*	WM03.230	Tissue from SDA plate [a]	Phenol chloroform	125014	54061	51340
Aspergillus	*niger*	WM06.98	Tissue from SDA plate [a]	Zymo kit	171615	65065	59406
Aspergillus	sp.	OCL015	Tissue from PDA plate	Qiagen kit	249468	42899	39988
Blastobotrys	*proliferans*	WM07.12	Tissue from SDA plate [a]	Phenol chloroform	133835	39631	37315
Candida	*albicans*	WM229	Tissue from SDA plate [a]	Zymo kit	91031	26114	25597
Candida	*metapsilosis*	WM01.56	Tissue from SDA plate [a]	Zymo kit	127633	40257	37426
Candida	*orthopsilosis*	WM03.414	Tissue from SDA plate [a]	Zymo kit	104214	32490	29765
Candida	*parapsilosis*	WM02.200	Tissue from SDA plate [a]	Zymo kit	135720	45958	42338
Candida	sp.	WM28	Tissue from SDA plate [a]	Zymo kit	109905	38085	35120
Cladophialophora	sp.	CLM599	Tissue from PDA plate [b]	Qiagen kit	115477	28063	25917
Clavispora	*lusitaniae*	WM18	Tissue from SDA plate [a]	Zymo kit	352768	141856	131936
Cortinarius	*globuliformis*	CM4	Fruiting tissue	Qiagen kit	347423	128993	117090
Cryptococcus	*zero*	OCL040	Tissue from PDA plate [b]	Qiagen kit	167818	42373	39235
Debaryomyces	sp.	WM03.458	Tissue from SDA plate [a]	Phenol chloroform	174974	35837	33499
Diaporthe	*foeniculina*	OCL060	Tissue from PDA plate [b]	Qiagen kit	206161	42329	39836
Diaporthe	sp.		Tissue from PDA plate [b]	Qiagen kit	198500	29833	27941
Discula	*quercina*	OCL067	Tissue from PDA plate [b]	Qiagen kit	172601	32847	30504
Discula	*quercina*	OCL068	Tissue from PDA plate [b]	Qiagen kit	188353	33438	31996
Dothiorella	*vidmadera*		Tissue from PDA plate [b]	Qiagen kit	204777	47318	44257
Entoleuca	sp.	OCL052	Tissue from PDA plate [b]	Qiagen kit	155158	33941	31356
Fusarium	*oxysporum*	Race3	Tissue from PDA plate [b]	Qiagen kit	382450	131411	123742
Geotrichum	*candidum*	WM17.23	Tissue from SDA plate [a]	Phenol chloroform	152933	8485	7805
Kluyveromyces	*marxianus*	WM13	Tissue from SDA plate [a]	Zymo kit	115282	31150	28382
Kluyveromyces	sp.	WM04.172	Tissue from SDA plate [a]	Zymo kit	370154	165113	152736
Kodamaea	*ohmeri*	WM10.200	Tissue from SDA plate [a]	Phenol chloroform	111257	38931	36478
Meyerozyma	*guilliermondii*	WM02.361	Tissue from SDA plate [a]	Phenol chloroform	211853	20333	18944
Penicillium	*chrysogenum*	WM06.341	Tissue from SDA plate [a]	Zymo kit	192173	78105	72307
Pichia	*kudriavzevii*	WM03.103	Tissue from SDA plate [a]	Zymo kit	122601	35604	33244
Pichia	*membranifaciens*	WM324	Tissue from SDA plate [a]	Zymo kit	104844	29540	26937
Puccinia	*striiformis-tritici*	104E	Fungal spores	Phenol chloroform	272465	122080	113337
Pyrenophora	*tritici-repentis*	Ptr8814	Tissue from PDA plate [b]	Qiagen kit	260896	97584	90015
Quambalaria	*cyanescens*	OCL055	Tissue from PDA plate [b]	Qiagen kit	205404	49780	46171
Rhodotorula	*mucilaginosa*	WM09.204	Tissue from SDA plate [a]	Zymo kit	318405	127801	117801
Saccharomyces	*cerevisiae*	YH2Gold	Tissue from YPD media [c]	Qiagen kit	96837	33025	30260
Scedosporium	*boydii*	WM09.122	Tissue from SDA plate [a]	Zymo kit	331947	102481	93723
Tapesia	*yallundae*	OCL029	Tissue from PDB	Qiagen kit	223186	59651	55589
Tapesia	*yallundae*	OCL031	Tissue from PDB	Qiagen kit	213143	52944	49481
Tuber	*brumale*		Fruiting tissue	Qiagen kit	275035	80614	74232
Wickerhamomyces	*anomalus*	WM03.505	Tissue from SDA plate [a]	Phenol chloroform	193187	45720	42589
Yamadazyma	*mexicana*	WM805	Tissue from SDA plate [a]	Phenol chloroform	179240	45093	42369
Yamadazyma	*scolyti*	WM06.835	Tissue from SDA plate [a]	Phenol chloroform	136650	37159	34841
Yarrowia	*lipolytica*	WM599	Tissue from SDA plate [a]	Phenol chloroform	141238	35950	33873
Zygoascus	*hellenicus*	WM02.460	Tissue from SDA plate [a]	Phenol chloroform	229073	36666	34002
Zymoseptoria	*tritici*	WA332	Tissue from PDA plate [b]	Qiagen kit	413127	143363	133089

Chapter 6
Nanopore Sequencing of Clinical Samples Using MinION Flow Cells

6.1 Long-read sequencing based clinical metagenomics for the detection and confirmation of *Pneumocystis jirovecii* directly from clinical specimens: A paradigm shift in mycological diagnostics

This manuscript was published in Medical Mycology in July 2020 (volume 58, issue 5, pages 650-660), published online as of the 23rd of November 2019. Hoang M.T.V. was involved in the acquisition and analysis of data, the design of the research and preparing the manuscript. Irinyi L. is the primary author and contributed to all aspects of this study. Pasic L. was involved in the acquisition of data, the design of the research and preparing the manuscript. Hu Y., Rathjen J., Stone E., and Schwessinger B. were involved in the data analysis. Basu I., McKinney W., and Morris A.J. were involved in sample acquisition. Meyer W. was involved in designing the research, project administration and funding acquisition. All authors were involved in the revision of the manuscript and support its inclusion in this thesis.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Minh Thuy Vi Hoang
30th June 2023

As a supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Professor Wieland Meyer
30th June 2023

ISHAM INTERNATIONAL SOCIETY FOR HUMAN AND ANIMAL MYCOLOGY

Medical Mycology, 2020, 58, 650–660
doi:10.1093/mmy/myz109
Advance Access Publication Date: 23 November 2019
Original Article

Original Article

Long-read sequencing based clinical metagenomics for the detection and confirmation of *Pneumocystis jirovecii* directly from clinical specimens: A paradigm shift in mycological diagnostics

Laszlo Irinyi[1,2,3,†], Yiheng Hu[6,†], Minh Thuy Vi Hoang[1,2,3], Lana Pasic[1,2,3], Catriona Halliday[4], Menuk Jayawardena[4], Indira Basu[8], Wendy McKinney[8], Arthur J. Morris[8], John Rathjen[6], Eric Stone[6,7], Sharon Chen[1,2,4], Tania C. Sorrell[1,2], Benjamin Schwessinger[6,*] and Wieland Meyer[1,2,3,5,*]

[1]Molecular Mycology Research Laboratory, Centre for Infectious Diseases and Microbiology, Faculty of Medicine and Health, Sydney Medical School, Westmead Clinical School, The University of Sydney, Sydney, NSW, Australia, [2]Marie Bashir Institute for Infectious Diseases and Biosecurity, The University of Sydney, Sydney, NSW, Australia, [3]Westmead Institute for Medical Research, Westmead, NSW Australia, [4]Centre for Infectious Diseases and Microbiology-Laboratory Services, Institute for Clinical Pathology and Medical Research, NSW Health Pathology, Westmead, NSW, Australia, [5]Westmead Hospital (Research and Education Network), Westmead, NSW, Australia, [6]Research School of Biology, Australian National University, Canberra, ACT, Australia, [7]ANU-CSIRO Centre for Genomics, Metabolomics and Bioinformatics, Canberra, ACT, Australia and [8]Microbiology Department, LabPLUS, Auckland City Hospital, Auckland, New Zealand

*To whom correspondence should be addressed. Prof. Wieland Meyer, PhD, Molecular Mycology Research Laboratory, CIDM, Westmead Institute for Medical Research, 176. Hawkesbury Road, Westmead, NSW 2145, Australia, Tel: +61 2 86273430, E-mail: wieland.meyer@sydney.edu.au and, Benjamin Schwessinger, PhD, Division of Plant Science, Research School of Biology, College of Science, Linnaeus Building (134), Linnaeus Way The Australian National University, Canberra ACT 0200 Australia, Tel: +61 2 61257794, E-mail: benjamin.schwessinger@anu.edu.au

†L.I., and Y.H. are equally contributed to this manuscript.

Received 18 July 2019; Revised 26 August 2019; Accepted 19 November 2019; Editorial Decision 25 September 2019

Abstract

The advent of next generation sequencing technologies has enabled the characterization of the genetic content of entire communities of organisms, including those in clinical specimens, without prior culturing. The MinION from Oxford Nanopore Technologies offers real-time, direct sequencing of long DNA fragments directly from clinical samples. The aim of this study was to assess the ability of unbiased, genome-wide, long-read, shotgun sequencing using MinION to identify *Pneumocystis jirovecii* directly from respiratory tract specimens and to characterize the associated mycobiome. *Pneumocystis* pneumonia (PCP) is a life-threatening fungal disease caused by *P. jirovecii*. Currently, the diagnosis of PCP relies on direct microscopic or real-time quantitative polymerase chain reaction (PCR) examination of respiratory tract specimens, as *P. jirovecii* cannot be cultured readily *in vitro*. *P. jirovecii* DNA was detected in bronchoalveolar lavage (BAL) and induced sputum (IS) samples from three patients with confirmed PCP. Other fungi present in the associated mycobiome included known human pathogens (*Aspergillus, Cryptococcus, Pichia*) as well as commensal species (*Candida, Malassezia, Bipolaris*). We have established optimized sample preparation conditions for the generation of high-quality data, curated databases, and data analysis tools, which are key to the application of long-read MinION sequencing leading to a fundamental new approach in fungal diagnostics.

Key words: mycoses, molecular diagnostics, clinical metagenomics, MinION, *Pneumocystis*.

© The Author(s) 2019. Published by Oxford University Press on behalf of The International Society for Human and Animal Mycology. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Downloaded from https://academic.oup.com/mmy/article/58/5/650/5638042 by University of Sydney user on 04 May 2021

Introduction

Rapid, reliable, and accurate detection of fungal pathogens enables early initiation of appropriate antifungal therapy and improved clinical outcomes of invasive fungal diseases. Conventional diagnostic methods have limitations, such as low sensitivity and/or specificity, long turnaround times, are time consuming, and have a restricted ability to detect multiple pathogens simultaneously. Next generation sequencing (NGS) is a powerful, culture-independent tool that can characterize mixed microbial communities.[1] The feasibility of direct fungal identification and characterization by NGS short-read sequencing has been demonstrated in human oral and respiratory mycobiome studies.[2,3] Third generation sequencing technologies, such as nanopore sequencing (Oxford Nanopore Technologies [ONT]), generate long sequence reads (kb to Mb), which enable more accurate mapping to reference genomes and real-time data analysis,[4] with superior discriminatory power in microbial identification.[5] Currently, the smallest sequencing platform offered by ONT is the MinION, a palm-sized sequencing device capable of generating ultra-long sequence reads in real-time, with low initial start-up costs and ease of use. It is uniquely suited for rapid diagnostics. The instrument has been applied successfully in bacterial outbreak investigation, real-time virus detection, and antibiotic resistance studies.[6] However, direct detection of pathogens in clinical specimens using a metagenomics approach is in its infancy and no application has previously been trialed for fungal pathogens.

The fungal microbiome (mycobiome) of bronchoalveolar lavage (BAL) and induced sputum (IS) is not well characterized, particularly in relation to infectious fungal diseases, such as candidiasis or *Pneumocystis* pneumonia (PCP). This is mainly because the respiratory tract was considered to be a sterile, harsh environment with a strong immune surveillance and relatively limited access to nutrients.[7] Until recent years, studies using short-read sequencing described complex microbiomes from the respiratory tract, including smokers and people with different respiratory diseases.[7,8] However, most of those studies only well characterized the bacterial compositions, while the fungal community remains largely unknown.

PCP, due to the opportunistic pathogen *Pneumocystis jirovecii*, has a high morbidity and mortality in patients infected with human immunodeficiency virus (HIV), organ transplant recipients, and patients with other causes of immunocompromise.[9] Diagnosis relies on identification of fungal cysts (or trophozoites) by direct microscopy on IS or BAL fluid using giemsa, toluidine blue or commercially available *Pneumocystis* specific immunofluorescent probes, but identification success varies and is largely dependent on the experience and skill of the laboratory personnel.[10,11] To overcome these issues, various polymerase chain reaction (PCR) based methods have been developed, such as standard or quantitative PCR usually targeting 400–700 bp long fragments.[12–16] Previous studies have shown that the sensitivity of PCR based methods can be improved by using a multicopy gene target, such as the mtLSU rRNA or the major surface glycoprotein,[14,17,18] or to use nested PCR to increase the detection rate.[19]

This study aimed to prove the utility of nanopore based metagenomics as part of the diagnosis of laboratory-confirmed cases of PCP from respiratory samples and to characterize the associated mycobiome of *P. jirovecii*.

Methods

Ethics

Ethical approval was not required for this as per institutional guidelines.

Sample collection and storage

We collected four induced sputum (IS) and two bronchoalveolar lavage (BAL) samples used for routine microbiological examination, including culture and/or *P. jirovecii* PCR. Three clinical samples were PCR positive for *Pneumocystis*, as per previous routine diagnosis (Pj_1, Pj_2, and Pj_3), and the other three were used as controls (C_1, C_2, and C_3) (Table 1).

Routine clinical microbiological testing

P. jirovecii DNA was detected using an "in house" real-time quantitative PCR assay that targets the *β-tubulin* or *surface glycoprotein* gene.[20]

DNA extraction for metagenomics analysis

Total DNA was extracted using the BAL DNA Isolation Kit (cat. no. 46 200, Norgen, Canada) following Irinyi et al.[21] adapted from the manufacturer's instructions. The detailed DNA extraction protocol can be found at: protocol.io (dx.doi.org/10.17504/protocols.io.6enhbde). Briefly, samples were ground in liquid nitrogen to break fungal cell walls before the commercial kit was used. To extract DNA, slurry D solution was added to liquefy the viscous sputum samples and mixed well. After a 5 minute low speed centrifuge step, Proteinase K and Lysozyme were added to the precipitated slurry pellet and incubated at 60°C for 20 minutes. After the incubation, the mixture was subjected to three washing steps with solutions supplied in the commercial Kit and ethanol. The DNA from the spin-column was eluted with 100 µl sterile water. The DNA concentration of the samples was determined using DeNovix ds-DNA Broad Range Fluorescent Assay Kit (Denovix, Wilmington, DE, USA). The average size of DNA was determined by agarose gel electrophoresis (1.5%) of 100 ng DNA for 1 hour at 100 volts.

Downloaded from https://academic.oup.com/mmy/article/58/5/650/5638042 by University of Sydney user on 04 May 2021

Table 1. Sample characteristics, summary of nanopore sequencing and taxonomy assignments obtained by using WIMP and BLAST for four induced sputum (IS) (C_1, C_2, C_3, Pj_1) and two bronchoalveolar lavage (BAL) specimens (Pj_2, Pj_3).

Sample	C_1		C_2		C_3		Pj_1		Pj_2		Pj_3	
Origin	CIDMLS, ICPMR, Westmead Hospital		CIDMLS, ICPMR, Westmead Hospital		CIDMLS, ICPMR, Westmead Hospital		Auckland City Hospital		CIDMLS, ICPMR, Westmead Hospital		CIDMLS, ICPMR, Westmead Hospital	
Specimen type	IS		IS		IS		IS		BAL		BAL	
Pneumocystis PCR result	Not Performed		Negative		Not Performed		Positive		Positive		Positive	
Total amount of DNA used in the DNA library preparation (ng)	100		100		100		100		33		33	
Total yield (Mbps)	2208		882		1599		49		388		320	
Mean length (bp)	626		1129		611		657		727		836	
No. of reads	3526,787		781,511		2616,428		81,956		535,411		383,693	
No. of passed QC reads*	3525,062 (99.95%)		781,160 (99.96%)		2615,142 (99.95%)		67,251 (82.06%)		499,125 (93.22%)		359,230 (93.62%)	
Workflow	WIMP	BLAST	WIMP	BLAST	WIMP	BLAST	WIMP	BLAST	WIMP	BLAST	WIMP	BLAST
No. of classified reads	3081,681 (87.42%)	3179,709 (90.2%)	747,688 (95.72%)	753,110 (96.41%)	2268,438 (86.74%)	2325,136 (88.91%)	27,500 (40.89%)	36,055 (53.61%)	467,525 (93.67%)	479,572 (96.08%)	329,200 (91.64%)	338,625 (94.26%)
No. of unclassified reads	443,381 (12.58%)	345,353 (9.8%)	33,472 (4.28%)	28,050 (3.59%)	346,704 (13.26%)	290,006 (11.09%)	39,751 (59.11%)	31,196 (46.39%)	31,600 (6.33%)	19,553 (3.92%)	30,030 (8.36%)	20,605 (5.74%)
No. of human reads (human/mammal)	2921,496 (82.88%)	2917,433/77,385 (82.76%/2.2%)	744,046 (95.25%)	734,343/11,133 (94.01%)/(1.43%)	2244,710 (85.84%)	2230,428/66,793 (85.29%)/(2.55%)	7010 (10.42%)	7099/137 (10.5%)/(0.2%)	342,960 (68.71%)	345,145/5956 (69.15%)/(1.19%)	318,312 (88.61%)	317,374/6120 (88.35%)/(1.7%)
No. of microbial reads	160,185 (4.54%)	184,891 (5.23%)	3642 (0.47%)	7634 (0.98%)	23,728 (0.91%)	27,915 (1.07%)	20,490 (30.47%)	28,819 (42.85%)	124,565 (24.96%)	128,471 (25.74%)	10,888 (3.03%)	15,131 (4.21%)
No. of bacterial reads	139,749 (3.96%)	156,531 (4.44%)	872 (0.11%)	1446 (0.19%)	5189 (0.2%)	7971 (0.3%)	20,274 (30.15%)	25,495 (37.91%)	123,195 (24.68%)	126,793 (25.4%)	9939 (2.77%)	12,437 (3.46%)
No. of fungal reads	9583 (0.27%)	73 (0.002%)	1315 (0.17%)	18 (0.0023%)	8847 (0.34%)	77 (0.003%)	124 (0.18%)	57 (0.08%)	589 (0.12%)	69 (0.01%)	532 (0.15%)	6 (0.0016%)
No. of *Pneumocystis* reads	11	0	2	0	19	1	14	14	9	9	1	2

Downloaded from https://academic.oup.com/mmy/article/58/5/650/5638042 by University of Sydney user on 04 May 2021

Library preparation

DNA sequencing libraries were prepared using the Ligation Sequencing 1D SQK-LSK108 and Native Barcoding Expansion (PCR-free) EXP-NBD103 Kits from ONT (UK), as adapted by Hu and Schwessinger,[22] which was adapted from the manufacturer's instructions with the omission of DNA fragmentation and DNA repair. DNA was first cleaned up using a 1× volume of Agencourt AMPure XP beads (cat. no. A63881, Beckman Coulter, Indianapolis, IN, USA), incubated at room temperature with gentle mixing for 5 minutes, washed twice with 200 μl fresh 70% ethanol, the pellet was allowed to dry for 2 minutes, and the DNA was eluted in 51 μl nuclease free water and quantified using NanoDrop® (Thermo Fisher Scientific, Waltham, MA, USA) and Promega Quantus™ Fluorometer (cat. No. E6150, Promega, Madison. WI, USA) follow the manufacturer's instructions. All DNA samples showed an absorbance ratio of A260/A280 >1.8 and an A260/A230 ratio of >2.0 from the NanoDrop®. DNA was end-repaired using NEBNext Ultra II End-Repair/dA-tailing Module (cat. No. E7546, New England Biolabs [NEB], Ipswich, MA, USA) by adding 7 μl Ultra II End-Prep buffer, 3 μl Ultra II End-Prep enzyme mix. The mixture was incubated at 20°C for 10 minutes and 65°C for 10 minutes. A 1× volume (60 μl) Agencourt AMPure XP clean-up was performed and the DNA was eluted in 31 μl nuclease free water. Barcoding reaction was performed by adding 2 μl of each native barcode and 20 μl NEB Blunt/TA Master Mix (cat. No. M0367, New England Biolabs [NEB], USA) into 18 μl DNA, mixing gently and incubating at room temperature for 10 minutes. A 1× volume (40 μl) Agencourt AMPure XP clean-up was then performed and the DNA was eluted in 15 μl nuclease free water. Ligation was then performed by adding 20 μl Barcode Adapter Mix (EXP-NBD103 Native Barcoding Expansion Kit, ONT, UK), 20 μl NEBNext Quick Ligation Reaction Buffer, and Quick T4 DNA Ligase (cat. No. E6056, New England Biolabs [(NEB], USA) to the 50 μl pooled equimolar barcoded DNA, mixing gently and incubating at room temperature for 10 minutes. The adapter-ligated DNA was cleaned-up by adding a 0.4× volume (40 μl) of Agencourt AMPure XP beads, incubating for 5 minutes at room temperature and resuspending the pellet twice in 140 μl ABB provided in the SQK-LSK108 kit. The purified-ligated DNA was resuspended by adding 15 μl ELB provided in the SQK-LSK108 (ONT, UK) kit and resuspending the beads. The beads were pelleted again and the supernatant transferred to a new 0.5 ml DNA LoBind tube (cat. no. 00 301 22348, Eppendorf, Germany).

Sequencing

In total, three independent sequencing reactions were performed on a MinION flow cell (R9.4, ONT) connected to a MK1B device (ONT) operated by the MinKNOW software (version 2.0.2): one reaction each for sample Pj_1, the control group and samples Pj_2 and Pj_3. Each flow cell was primed with 1 ml of priming buffer comprising 480 μl Running Buffer Fuel Mix (RBF, ONT) and 520 μl nuclease free water. In total, 12 μl of amplicon library was added to a loading mix including 35 μl RBF, 25.5 μl Library Loading beads (ONT library loading bead kit EXP-LLB001, batch number EB01.10.0012) and 2.5 μl water with a final volume of 75 μl and then added to the flow cell via the SpotON sample port. The 'NC_48Hr_sequencing_FLO-MIN106_SQK-LSK108' protocol was executed through MinKNOW after loading the library, and run for 48 hours. Raw fast5 files were processed using Albacore 2.3.1 software (ONT) for basecalling, barcode de-multiplexing and quality filtering (Phred quality (Q) score of >7) as per the manufacturer's recommendations. Basecalled fastq files were uploaded into NCBI Short Reads Archive under the Bioproject number PRJNA524377 including six accession numbers: SRX5439674, SRX5439673, SRX5439672, SRX5439671, SRX5439670, and SRX5439669.

Data analysis

All reads that passed Albacore quality filtering and recognized multiplexed barcodes were treated in parallel as follows: First, barcode and adapter sequences were trimmed from the ends of reads using Porechop.[23] To identify middle adapter sequences by Porechop, a 95% threshold was set, and reads with middle adapter sequences were discarded. Secondly, seqtk (https://github.com/lh3/seqtk) was used to convert the processed fastq file into fasta format. The 'What's in my pot' (WIMP) quantitative online analysis tool[24] and nucleotide BLAST (blastn, Basic Local Alignment Search Tool)[25] analyses were then performed independently. BLAST searches were performed using the fasta files for each barcode as queries against the National Center for Biotechnology Information (NCBI) nucleotide (nt) database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/, downloaded 09/2018). Individual reads were assigned to a specific DNA sequence based on their best blastn hit (E-value <0.01).

After PoreChop, WIMP uses fastq files directly through EPI2ME Desktop Agent (vers. v2.58). A Q score of ≥7 was set in the beginning of the workflow. Fastq files from each barcode were separated into files with 4000 reads within a folder using fastq-splitter.pl script (http://kirill-kryukov.com/study/tools/fastq-splitter/) before uploading onto the WIMP server through the EPI2ME Desktop Agent. The WIMP analysis output file and quality control csv file for each sample were deposited at Figshare https://doi.org/10.6084/m9.figshare.7772852.v1.

We used Python programming language to summarize relevant information from the sequencing summary files and output files from WIMP and BLAST. To facilitate the reproducibility of our results, all python modules, programming environment, and three scripts used for all analysis and figure plotting were cited and deposited at https://github.com/Yiheng323/clinical_diagnosis_of_pneumocystis.git.

Downloaded from https://academic.oup.com/mmy/article/58/5/650/5638042 by University of Sydney user on 04 May 2021

Results

Metagenomic sequencing of BAL and IS samples

DNA extraction from IS and BAL samples yielded various amounts of dsDNA per sample (33 ng from BAL samples and hundreds of nanograms from IS samples). The total amount of DNA we used in the library preparation was 33–100 ng for each sample (Table 1). After the sequencing, we used Albacore software to convert raw sequencing signal to nucleotide sequence. Based on read mean Q scores, Albacore grouped reads into pass (Q score >7) or fail (Q score ≤7). We only proceeded with analysis of sequencing reads with a Q score above 7 and reads with barcode sequences identified by Albacore at both ends (see Table 1). The Pj_1 IS sample yielded only 49 Mbps of total with 67,251 reads passing the quality control. The other two *Pneumocystis* positive BAL samples resulted in 388 and 320 Mbps of total sequence length with 499,125 and 359,230 reads and a Q score >7, respectively. Sequencing of the control sample C_1 generated 2208 Mbps total sequences, representing the highest yield from all samples, followed by 1599 Mbps from sample C_3 and 882 Mbps from sample C_2 (Table 1).

Independent WIMP and BLAST taxonomic classification of *P. jirovecii*

To evaluate different analysis and diagnostic strategies we performed WIMP and BLAST independently on all quality filtered sequence data. We assigned the origin of each read based on a 'winner-takes-all' strategy, which considers the best match of each sequencing read defined by the lowest BLAST E-values or the highest WIMP scores. WIMP and BLAST classified 87–95% of reads in each sample, except for the sample Pj_1 where only 40.89% and 53.61% were classified, respectively.

Reads assigned to human (*Homo sapiens*) (or other mammalian species in case of the BLAST) ranged from 70% to 95% of total reads per sample, except for Pj_1 (~10%) (Table 1). The relative high error rate of nanopore sequencing may lead to misclassifications when comparing reads with reference genomes originated from closely related species. Therefore, we choose to examine the fungal population classified by WIMP and BLAST at genus level to tolerate misclassifications within the same genus. From the WIMP classification, the total fungal sequence length comprised 0.10% to 0.26% of the total sequence length of each group. When separating reads assigned to fungal genomes, the proportion of each genus was consistent (Fig. 1). The genus *Malassezia* ranked the highest in all samples, except sample Pj_1, where *Pneumocystis* reads were ranked the highest. Most of the fungal genera identified by WIMP were shared among all samples, such as *Aspergillus* (3.42%–5.28%), *Phycomyces* (3.80%–7.09%), and *Lobosporangium* (3.23%–6.11%). Unexpectedly, *Pneumocystis* reads were detected in both pneumocystis positive and control samples using the WIMP algorithm with an abundance of total fungal sequence length at 0.49%–14.42% and 0.13% – 0.23%, respectively. Although the abundance of *Pneumocystis* in the positive group was higher than that in the *Pneumocystis* control group, this result suggests that the WIMP classification alone does not provide enough confidence for diagnosis. Compared to WIMP, BLAST analysis against the NCBI nucleotide database classified much fewer reads as fungi (Table 1), comprising slightly over 0.1% in the three *Pneumocystis* positive samples and around 0.001% in all the rest of the samples. *Pneumocystis* reads were mostly detected in the *Pneumocystis* positive group, although one read from sample C_3 was assigned to the *P. jirovecii* genome.

While BLAST analysis resulted in less reads classified into fungi, it provided a more close-to-reality diagnosis than the WIMP analysis. However, it is worth noting that except for sample Pj_3, the total sequence length of *Pneumocystis* reads classified by either WIMP or BLAST never ranked first among all fungal reads in any *Pneumocystis* positive sample.

Improved diagnosis of *P. jirovecii* by cross-checking fungal reads identified by WIMP with BLAST classification

Many of the fungal genera to which reads were assigned were plant-associated fungi. Examples include *Trichoderma*, *Colletotrichum*, and *Endocarpon* from the WIMP classification and *Rhizophagus* from the BLAST classification (Fig. 1). It is probable that these represent false positives, since no previous study reported their presence in human IS or BAL samples. To improve the specificity of *P. jirovecii* identification by WIMP or BLAST, we compared their concordant classification result for each sample at different taxonomic ranks. Interestingly, even at the kingdom level, the percentage of concordant reads was very low (Fig. 2A), around 10% of the fungal sequences were concordant in two samples and 0.15%–0.45% in the rest. To verify that this method of comparison was unbiased, we tested if such little overlap by different analysis methods was specific to fungi. Therefore, we performed the same comparison on bacterial reads between BLAST and WIMP. Unlike the result from the fungal classification, the general proportion of total sequence length of bacteria reads by WIMP was similar to the proportion assigned by BLAST within each sample (Table 1), and the majority of classifications were concordant for each read at kingdom level (Fig. 2B).

To further investigate the accordance of the read assignment we extracted the accordant fungal reads of each sample (Table 2). Both BLAST and WIMP analysis identified the same 24 reads as *P. jirovecii* within the *Pneumocystis* positive group. In contrast, no reads were assigned concordantly to *Pneumocystis* in the control group. Both BLAST and WIMP assigned 86 reads to 26 other fungal genera in addition to *P. jirovecii* including *Malassezia*, *Candida*, and *Cryptococcus* (Table 2). Overall, 98

Downloaded from https://academic.oup.com/mmy/article/58/5/650/5638042 by University of Sydney user on 04 May 2021

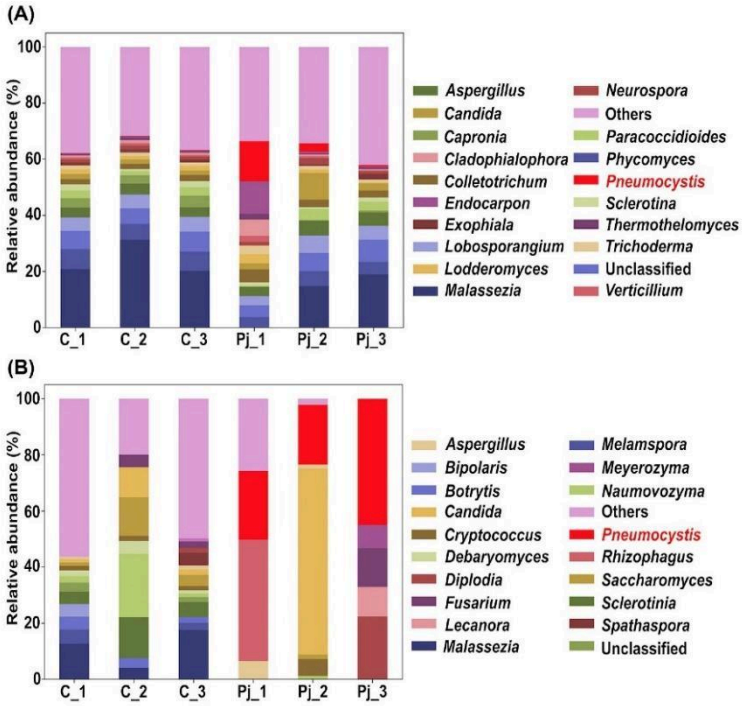


Figure 1. The proportion of total sequence length for 18 most common fungal genera identified by WIMP (**A**) and BLAST (**B**). Columns represent four induced sputum (IS) (C_1, C_2, C_3, Pj_1) and two bronchoalveolar lavage (BAL) specimens (Pj_2, Pj_3). The *y*-axis indicates relative abundance of total sequence lengths for each genus represented by each color. Reads without classified genera but assigned to fungal kingdom were labeled as 'Unclassified.' Fungal genera other than the most common 18 genera were grouped as 'Others.' Genus *Pneumocystis* was labeled with red color.

reads from the concordant 110 fungal reads were assigned to the same species by BLAST and WIMP.

Discussion

The main aim of this study was to apply long-read sequencing to identify human pathogenic fungal species, for example, *P. jirovecii*, directly from clinical samples, such as BAL and IS. We aimed to test if the results obtained by routine diagnostic methods can be confirmed by direct metagenomic shotgun long-read sequencing in order to evaluate its potential as future diagnostic tool for human fungal pathogens. As a result of this study, we propose a nanopore based metagenomics workflow to characterize human fungal pathogens from sample collection to data interpretation (Fig. 3).

MinION based long-read sequencing has been used successfully in a clinical setting to identify Chikungunya, Ebola, and Hepatitis C virus from human blood samples[26] or *Mycobacterium tuberculosis* from direct respiratory samples.[27] In this study, we extended its application to identify fungal pathogens from human respiratory samples, such as *P. jirovecii*, to verify the traditional diagnosis of PCP from a clinical specimen and explored the potential mycobiome of this pathogen.

In addition to *P. jirovecii*, many fungal taxa were detected in the herein studied nonsterile *P. jirovecii* positive BAL and IS specimens. Both the BLAST and WIMP algorithms classified reads into fungal taxa other than *Pneumocystis*, indicating that the disease-causing organism might not necessarily be the most abundant mycobiome member. Without adequate training and additional patient-related disease background data this may lead to the wrong identification of the disease-causing agent. Many of these taxa are part of the human respiratory mycobiome in healthy individuals,[3] such as previously described *Aspergillus* spp., *Candida* spp., *Lobosporangium* spp., and *Malassezia* spp.[28] *Malassezia* spp. are the main member of the microbiome of healthy skin,[29] and their presence in our experiment are likely due to contamination during the sampling process.

Downloaded from https://academic.oup.com/mmy/article/58/5/650/5638042 by University of Sydney user on 04 May 2021

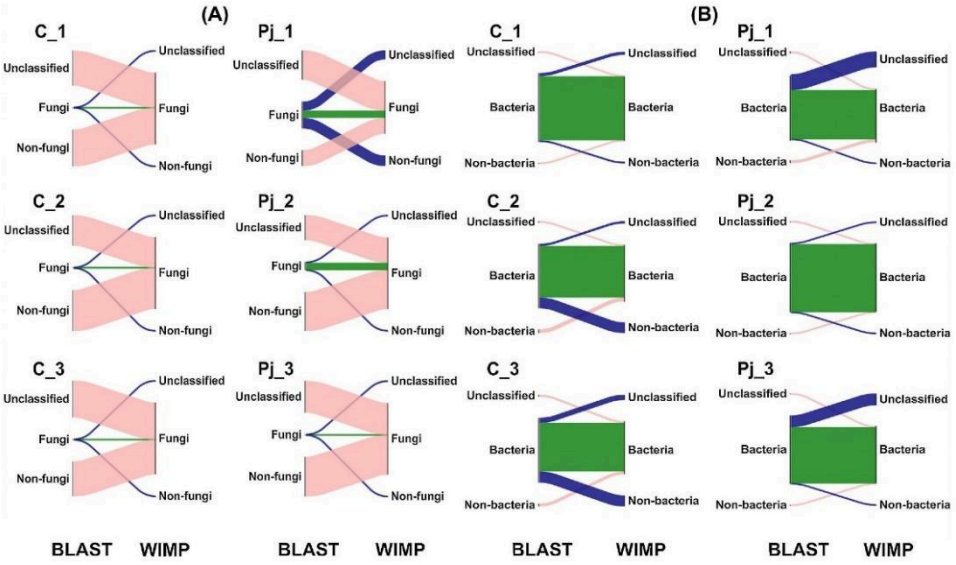


Figure 2. Comparison of fungal reads (**A**) and bacterial reads (**B**) classified by WIMP and BLAST from four induced sputum (IS) (C_1, C_2, C_3, Pj_1) and two bronchoalveolar lavage (BAL) specimens (Pj_2, Pj_3). Green lines indicate species detection concordance, showing a much higher precision for bacteria than fungi, pink lines indicate WIMP and blue lines BLAST specific identifications.

However, some of the genera, such as *Paracoccidioides* spp. identified by the WIMP algorithm herein, are clearly false positives. Species in the genus *Paracoccidioides* can cause systemic mycoses but they are geographically restricted to subtropical areas of Latin America.[30] Therefore, the initial detection of *Paracoccidioides* spp. by the WIMP workflow in our study was alarming. *Paracoccidioides* reads identified by WIMP were then subjected to BLAST search against the NCBI database and none of them were identified as *Paracoccidioides* spp. After crosschecking the nonconcordant reads, such as those corresponding to *Paracoccidioides* spp. and other plant pathogenic fungal species classified by WIMP, many were unclassified even at kingdom level when using BLAST. In the *Pneumocystis* positive group, all *Pneumocystis* reads along other human pathogens or commensal species, such as *Candida* and *Cryptococcus* were concordantly classified by both WIMP and BLAST. We compared all fungal, Chordata, and bacterial reads classified by WIMP and BLAST to see whether they are concordant. We found that bacterial and Chordata reads are largely concordant, while the fungal reads are not. It is not clear why the level of concordance in the classification of bacteria and Chordata was high but in the classification of fungi was low. One of the reasons might be the low number of fungal species present in the genome database of NCBI compared to bacteria. These results indicated that false-positives and error-prone reads represent a challenge for metagenomics studies, especially at the species level.[31,32] Taxonomic profiling and binning programs proved to be accurate at high taxonomic ranks, but their performance decreased significantly below family level.[32] Another reason might be that fungal genomes are usually poorly annotated and more prone to cross-species genome contamination due to the complex structure of eukaryotic genomes, which may result in incorrect taxonomic binning.[31–35]

The results show that the BLAST analysis delivers a more accurate classification than WIMP. BLAST is the most widely used alignment-based sequence similarity search algorithm.[36] It consists of aligning a query sequence to a sequence database to find the best statistically significant matches, coming with major disadvantages, being generally memory and time-consuming, limiting its use for metagenome-scale sequence data analysis. To overcome these shortcomings, alignment free algorithms have been suggested.[37] In recent years, k-mer-based alignment free metagenomics analysis and comparison methods, for example, Centrifuge (used by WIMP)[35] and Kraken[38] have become popular, providing much faster results.

In our case, the combination of BLAST and WIMP analysis provides a strong confirmation of the *P. jirovecii* read assignment in the *Pneumocystis* positive group and overcomes the false positive classification of *Pneumocystis* in the control group, making a successful diagnosis of PCP feasible. Based on this finding, we suggest a combination of the two approaches in a clinical setting when looking at fungal identification. This enables in the first instance a fast and in the second instance accurate identification,

Downloaded from https://academic.oup.com/mmy/article/58/5/650/5638042 by University of Sydney user on 04 May 2021

Table 2. Number of reads classified at species level as fungi by WIMP and BLAST and their concordance in four induced sputum (IS) (C_1, C_2, C_3, Pj_1) and two bronchoalveolar lavage (BAL) specimens (Pj_2, Pj_3).

	C_1			C_2				C_3			
Species	WIMP	BLAST	Concordance	Species	WIMP	BLAST	Concordance	Species	WIMP	BLAST	Concordance
Agaricus bisporus	1	1	1	*Candida albicans*	1	1	1	*Agaricus bisporus*	1	1	1
Aspergillus oryzae	0	1	0	*Malassezia globosa*	2	1	1	*Arthrobotrys oligospora*	1	1	1
Bipolaris maydis	1	1	1	*Pichia kudriavzevii*	1	1	1	*Ascoidea rubescens*	1	1	1
Cryptococcus amylolentus	0	1	0	*Saccharomyces cerevisiae*	1	1	1	*Candida albicans*	0	1	0
Cryptococcus neoformans	0	1	0	*Sclerotinia sclerotiorum*	0	1	0	*Candida tropicalis*	1	0	0
Exserohilum turcicum	1	0	0					*Eremothecium gossypii*	1	1	1
Malassezia globose	4	3	3					*Exophiala mesophila*	1	1	1
Paracoccidioides lutzii	0	1	0					*Kluyveromyces marxianus*	0	1	0
Penicillium rubens	0	1	0					*Lodderomyces elongisporus*	1	0	0
Pseudocercospora fijiensis	0	1	0					*Malassezia globosa*	5	5	4
Sclerotinia sclerotiorum	1	1	1					*Neurospora crassa*	0	1	0
Sugiyamaella lignohabitans	1	0	0					*Sclerotinia sclerotiorum*	1	0	0
Tsuchiyaea wingfieldii	1	0	0					*Spathaspora passalidarum*	1	1	1
Unclassified Fungi	1	0	0					*Trichoderma reesei*	1	1	1
Vanderwaltozyma polyspora	1	0	0					*Trichosporon asahii*	0	1	0
								Unclassified Fungi	1	0	0
Total	12	12	6	Total	5	5	4	Total	16	16	10
Concordance %		50		Concordance %		80		Concordance %		62.5	
	Pj_1				Pj_2				Pj_3		
Species	WIMP	BLAST	Concordance	Species	WIMP	BLAST	Concordance	Species	WIMP	BLAST	Concordance
Coniophora puteana	1	1	1	*Candida auris*	1	1	1	*Meyerozyma guilliermondii*	1	1	1
Leptosphaeria biglobosa	0	1	0	*Candida dubliniensis*	45	45	45	*Pneumocystis jirovecii*	1	1	1
Pneumocystis jirovecii	14	14	14	*Cryptococcus neoformans*	5	5	5				
Thielavia terrestris	1	1	1	*Pneumocystis jirovecii*	9	9	9				
Unclassified Fungi	1	0	0								
Total	17	17	16	Total	60	60	60	Total	2	2	2
Concordance %		94		Concordance %		100		Concordance %		100	

Downloaded from https://academic.oup.com/mmy/article/58/5/650/5638042 by University of Sydney user on 04 May 2021

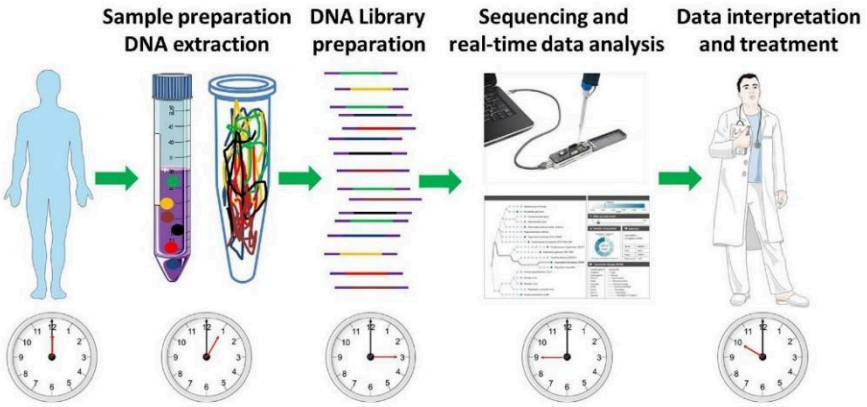


Figure 3. Schematic presentation of the nanopore based metagenomics pipeline to detect and characterize of human pathogens from sample collection to data interpretation.

by running first WIMP to filter out the human reads followed by BLAST to accurately classify the microbial reads.

Another major hurdle to overcome is the low proportion of pathogen DNA in comparison to the host DNA in the sample. This low pathogen DNA content makes it difficult to generate enough target species reads for downstream analyses. Our study is in agreement with previous studies showing that ~95% of raw NGS reads were of human DNA origin.[39] While there are bioinformatic tools to filter out human sequences from a large NGS data set, it still consumed most of the sequencing resources, lowering the sequencing depth of pathogens and the level of cost-effectiveness for pathogen diagnosis. Solutions to this problem can be to specifically lyse human cells and remove the human DNA prior to sequencing,[27,40] or to use microbial DNA enrichment methods to improve the pathogen:host DNA ratio. Both methods have been successfully employed in many studies to enrich bacterial DNA from clinical samples improving the pathogen:host read ratios.[41,42] However, in fungi previous enrichment studies focused on improving the sensitivity of PCR based diagnostics with limited success, and that's why this has rarely been applied to metagenomics studies.[41,43–45]

The success of any DNA based detection method depends on the lysis of the fungal cells in the clinical specimen. Fungi have thick cell walls that impede lysis and the recovery of nucleic acids, such as the *Pneumocystis* cystic form with hard-to-break-up cell walls.[46] Fungal DNA extraction protocols commonly require physical force to rupture the thick cell wall, which has the risk to fragment the DNA. It is crucially important to extract DNA of adequate quality, with a high purity, high molecular weight (HMW), and high concentration, to maximize detection sensitivity and specificity of long-read metagenomics. Therefore, the development of better DNA extraction methods for fungal DNA from complex tissues is a very important research question in regards to improving the accuracy of DNA based fungal pathogen detection.

Another bottleneck of metagenomic shotgun long-read sequencing is the lack of high-quality, curated reference databases of fungal whole genomes. Those databases need to be curated and highly accurate in terms of sequences, metadata, and phylogenetic coverage. False positives and irrelevant organisms, such as *Paracoccidioides* spp. detected in this study are partially due to the incorrect taxonomic identification and annotation of sequences present in the databases.[47]

In conclusion, we provided a proof-of-principle that metagenomic shotgun long-read sequencing with nanopore can identify the presence of a specific disease causing human pathogenic fungus, in our case *P. jirovecii*, directly from a clinical specimen. Yet many challenges (DNA quality, low DNA pathogen:host ratio, lack of quality-controlled reference databases, and bioinformatics tools) are yet to be overcome before a routine application in clinical settings. Despite all the challenges, we anticipate that within the next decade, shotgun metagenomics-based detection and characterization of fungal pathogens will be implemented and translated into routine diagnostic in public health laboratories.

Study limitations

The small sample size in this study is a limitation to detect any specific significant pathogen association or correlation with diseases status; therefore, strong conclusions in identifying pathogen associated sequence signals, other than the confirmation of the causative disease agent, cannot currently be drawn. The sole aim of this study was to apply long-read sequencing to identify fungal species directly from BAL and sputum induced samples.

Downloaded from https://academic.oup.com/mmy/article/58/5/650/5638042 by University of Sydney user on 04 May 2021

Acknowledgments

The authors acknowledge the Sydney Informatics Hub and the University of Sydney's high-performance computing cluster Artemis for providing the high-performance computing resources that have contributed to the research results reported within this paper. We thank Joey Lai Genomics Facility manager at the Westmead Institute for Medical Research for his assistance in DNA quantification. We thank Mari Miyamoto from Oxford Nanopore for her technical advice and interest in the project.

This study was supported by a National Health and Medical Research Council of Australia (NH&MRC) grant [no. APP1121936] to W.M., S.C., T.C.S., and Western Sydney Local Health District Research & Education Network Research Grant Scheme to W.M., T.C.S., S.C., and L.I. B.S. is supported by an Australian Research Council Future Fellowship FT180100024, and Y.H., E.S., J.R., and B.S. are supported by The Hermon Slade Foundation grant HSF_17_04. T.C.S. is a Sydney Medical Foundation Fellow.

Declaration of interest

The authors report no conflicts of interest associated with the manuscript. The authors alone are responsible for the content and the writing of the paper.

References

1. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. *Trends Genet*. 2014; 30: 418–426.
2. Ghannoum MA, Jurevic RJ, Mukherjee PK et al. Characterization of the oral fungal microbiome (Mycobiome) in healthy individuals. *PLoS Pathog*. 2010; 6: e1000713.
3. Nguyen LDN, Viscogliosi E, Delhaes L. The lung mycobiome: an emerging field of the human respiratory microbiome. *Front Microbiol*. 2015; 6: 89.
4. Faria NR, Sabino EC, Nunes MRT, Alcantara LCJ, Loman NJ, Pybus OG. Mobile real-time surveillance of Zika virus in Brazil. *Genome Med*. 2016; 8: 97.
5. Benitez-Paez A, Sanz Y. Multi-locus and long amplicon sequencing approach to study microbial diversity at species level using the MinION portable nanopore sequencer. *GigaScience*. 2017; 6: 1–12.
6. Leggett RM, Clark MD. A world of opportunities with nanopore sequencing. *J Exp Bot*. 2017; 68: 5419–5429.
7. Belkaid Y, Hand TW. Role of the microbiota in immunity and inflammation. *Cell*. 2014; 157: 121–141.
8. Salisbury ML, Han MK, Dickson RP, Molyneaux PL. Microbiome in interstitial lung disease: from pathogenesis to treatment target. *Curr Opin Pulm Med*. 2017; 23: 404–410.
9. Maini R, Henderson KL, Sheridan EA et al. Increasing *Pneumocystis* pneumonia, England, UK, 2000–2010. *Emerg Infect Dis*. 2013; 19: 386–392.
10. Sun T, Chess Q, Tanenbaum B. Morphologic criteria for the identification of *Pneumocystis carinii* in Papanicolaou-stained preparations. *Acta Cytol*. 1986; 30: 80–82.
11. Procop GW, Haddad S, Quinn J et al. Detection of *Pneumocystis jiroveci* in respiratory specimens by four staining methods. *J Clin Microbiol*. 2004; 42: 3333–3335.
12. Larsen HH, Masur H, Kovacs JA et al. Development and evaluation of a quantitative, touch-down, real-time PCR assay for diagnosing *Pneumocystis carinii* pneumonia. *J Clin Microbiol*. 2002; 40: 490–494.
13. Robert-Gangneux F, Belaz S, Revest M et al. Diagnosis of *Pneumocystis jirovecii* pneumonia in immunocompromised patients by real-time PCR: a 4-year prospective study. *J Clin Microbiol*. 2014; 52: 3370–3376.
14. Wakefield AE, Pixley FJ, Banerji S et al. Detection of *Pneumocystis carinii* with DNA amplification. *Lancet*. 1990; 336: 451–453.
15. Robberts FJ, Liebowitz LD, Chalkley LJ. Polymerase chain reaction detection of *Pneumocystis jiroveci*: evaluation of 9 assays. *Diagn Microbiol Infect Dis*. 2007; 58: 385–392.
16. Fauchier T, Hasseine L, Gari-Toussaint M, Casanova V, Marty PM, Pomares C. Detection of *Pneumocystis jirovecii* by quantitative PCR to differentiate colonization and pneumonia in immunocompromised HIV-positive and HIV-negative patients. *J Clin Microbiol*. 2016; 54: 1487–1495.
17. Wakefield AE, Pixley FJ, Banerji S et al. Amplification of mitochondrial ribosomal RNA sequences from *Pneumocystis carinii* DNA of rat and human origin. *Mol Biochem Parasitol*. 1990; 43: 69–76.
18. Larsen HH, Huang L, Kovacs JA et al. A prospective, blinded study of quantitative touch-down polymerase chain reaction using oral-wash samples for diagnosis of *Pneumocystis* pneumonia in HIV-infected patients. *J Infect Di*. 2004; 189: 1679–1683.
19. Alvarez-Martinez MJ, Miro JM, Valls ME et al. Sensitivity and specificity of nested and real-time PCR for the detection of *Pneumocystis jiroveci* in clinical specimens. *Diagn Microbiol Infect Dis*. 2006; 56: 153–160.
20. Brancart F, Rodriguez-Villalobos H, Fonteyne PA, Peres-Bota D, Liesnard C. Quantitative TaqMan PCR for detection of *Pneumocystis jiroveci*. *J Microbiol Methods*. 2005; 61: 381–387.
21. Irinyi L, Pasic L, Meyer W. DNA extraction from sputum. 2019; https://www.protocols.io/view/dna-extraction-from-sputum-6enhbde.
22. Hu Y, Schwessinger B. Amplicon sequencing using MinION optimized from 1D native barcoding genomic DNA. 2018; 1–17. https://www.protocols.io/view/amplicon-sequencing-using-minion-optimized-from-1d-mhkc34w.
23. Wick RR, Judd LM, Gorrie CL, Holt KE. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb Genom*. 2017; 3: e000132–e000132.
24. Juul S, Izquierdo F, Hurst A et al. What's in my pot? Real-time species identification on the MinION. *bioRxiv*. 2015: https://doi.org/10.1101/030742.
25. Camacho C, Coulouris G, Avagyan V et al. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009; 10: 421.
26. Greninger AL, Naccache SN, Federman S et al. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Med*. 2015; 7: 1–13.
27. Charalampous T, Kay GL, Richardson H et al. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. *Nat Biotechnol*. 2019; 37: 783–792.
28. Hogan DA, Willger SD, Dolben EL et al. Analysis of lung microbiota in bronchoalveolar lavage, protected brush and sputum samples from subjects with mild-to-moderate cystic fibrosis lung disease. *PLoS One*. 2016; 11: e0149998.
29. Gaitanis G, Magiatis P, Hantschke M, Bassukas ID, Velegraki A. The *Malassezia* genus in skin and systemic diseases. *Clin Microbiol Rev*. 2012; 25: 106–141.
30. San-Blas G, Nino-Vega G, Iturriaga T. *Paracoccidioides brasiliensis* and paracoccidioidomycosis: molecular approaches to morphogenesis, diagnosis, epidemiology, taxonomy and genetics. *Med Mycol*. 2002; 40: 225–242.
31. Lindgreen S, Adair KL, Gardner PP. An evaluation of the accuracy and speed of metagenome analysis tools. *Sci Rep*. 2016; 6: 19233.
32. Sczyrba A, Hofmann P, Belmann P et al. Critical assessment of metagenome interpretation-a benchmark of metagenomics software. *Nat Methods*. 2017; 14: 1063–1071.
33. Merchant S, Wood DE, Salzberg SL. Unexpected cross-species contamination in genome sequencing projects. *PeerJ*. 2014; 2: e675.
34. Dadi TH, Renard BY, Wieler LH, Semmler T, Reinert K. SLIMM: species level identification of microorganisms from metagenomes. *PeerJ*. 2017; 5: e3138.
35. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res*. 2016; 26: 1721–1729.
36. Boratyn GM, Camacho C, Cooper PS et al. BLAST: a more efficient report with usability improvements. *Nucleic Acids Res*. 2013; 41: W29–33.
37. Vinga S, Almeida J. Alignment-free sequence comparison-a review. *Bioinformatics*. 2003; 19: 513–523.
38. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol*. 2014; 15: R46.
39. Bates M, O'Grady J, Maeurer M et al. Assessment of the Xpert MTB/RIF assay for diagnosis of tuberculosis with gastric lavage aspirates in children in sub-Saharan Africa: a prospective descriptive study. *Lancet Infect Dis*. 2013; 13: 36–42.
40. Hasan MR, Rawat A, Tang P et al. Depletion of human DNA in spiked clinical specimens for improvement of sensitivity of pathogen detection by next-generation sequencing. *J Clin Microbiol*. 2016; 54: 919–927.

Downloaded from https://academic.oup.com/mmy/article/58/5/650/5638042 by University of Sydney user on 04 May 2021

41. Feehery GR, Yigit E, Oyola SO et al. A method for selectively enriching microbial DNA from contaminating vertebrate host DNA. *PLoS One*. 2013; 8: e76096.
42. Loonen AJ, Jansz AR, Stalpers J, Wolffs PF, van den Brule AJ. An evaluation of three processing methods and the effect of reduced culture times for faster direct identification of pathogens from BacT/ALERT blood cultures by MALDI-TOF MS. *Eur J Clin Microbiol Infect Dis*. 2012; 31: 1575–1583.
43. Amorim-Vaz S, Sanglard D. Novel approaches for fungal transcriptomics from host samples. *Front Microbiol*. 2016; 6: 1571–1571.
44. Springer J, Loeffler J, Heinz W et al. Pathogen-specific DNA enrichment does not increase sensitivity of PCR for diagnosis of invasive aspergillosis in neutropenic patients. *J Clin Microbiol*. 2011; 49: 1267–1273.
45. Thoendel M, Jeraldo PR, Greenwood-Quaintance KE et al. Comparison of microbial DNA enrichment tools for metagenomic whole genome sequencing. *J Microbiol Methods*. 2016; 127: 141–145.
46. De Stefano JA, Cushion MT, Sleight RG, Walzer PD. Analysis of *Pneumocystis carinii* cyst wall. I. Evidence for an outer surface membrane. *J Protozool*. 1990; 37: 428–435.
47. Bidartondo MI. Preserving accuracy in GenBank. *Science*. 2008; 319: 1616.

Downloaded from https://academic.oup.com/mmy/article/58/5/650/5638042 by University of Sydney user on 04 May 2021

Chapter 7
Nanopore Sequencing of Clinical Samples using Flongle Flow Cells

7.1 Introduction

The current gold standard in fungal diagnostics is DNA barcoding, which achieves rapid and accurate pathogen identification, and requires less labour and expertise than traditional methods [1]. Although DNA barcoding is able to reliably achieve the identification of most pathogenic fungal species, the technique is limited in the amount of information generated. Species identification based on the dual DNA barcoding scheme for pathogenic fungal species is reliant on two loci, the ITS1/2 region and the *TEF1α* gene locus [2]. Additionally, DNA barcoding is limited in its capacity to identify coinfections as multiple amplicons will appear as mixed peaks in Sanger sequencing chromatograms, making recognition and diagnosis of coinfections difficult [3, 4]. As such, the implementation of new sequencing technologies generating sequences from different organisms simultaneously could be a new way to overcome the drawbacks of DNA barcoding, providing multiple pathogen information [3].

NGS has emerged and has greatly improved the throughput of sequencing technologies, by parallel sequencing of multiple DNA templates to generate large amounts of data. NGS is highly suited to be applied to fungal diagnostics as sequencing is culture-independent, reducing the time required for accurate identification and thus drastically improving patient outcomes [5]. The incorporation of a metagenomic approach to NGS allows for sequencing to be performed on all genomic material within a sample, thereby identifying all microbial species present and characterising multiple microbial profiles in complex communities such as those in patient samples [4]. Metagenomic sequencing of all genomic material in samples potentially provides further information beyond identification that can be applied to improve patient prognosis. Previously identified antimicrobial resistance markers may be sequenced and identified from metagenomic sequencing and thereby be used to inform on antimicrobial treatment and thus limiting the inappropriate use of antimicrobials to reduce the spread of

resistance. Additionally, established genetic markers may be sequenced to provide epidemiological information for outbreak tracing, such as loci being part of multi-locus sequencing typing schemes, or to further categorise species into subtypes [6]. Metagenomic sequencing is not reliant on prior amplification of genomic material as such as is the case in metabarcoding, and so would provide a more accurate reflection of proportions of microbial species within communities [7]. This could be used to inform on pathogen identification in cases where the microbe in highest abundance would be the likely pathogen as is the case in sterile sites such as blood and cerebrospinal fluid (CSF). Due to the large amounts of sequencing data generated, metagenomic sequencing and the computation requirements are generally more expensive than amplicon sequencing.

Third generation nanopore-based sequencing by ONT can generate ultra-long sequencing reads of up to 2 million base pairs long, dependent on the length of the genomic material and the sample quantity and quality [8]. Nanopore sequencing relies on biological nanopores embedded in solid state membranes within flow cells. DNA or RNA strands undergo library preparation and are then processed through nanopores. During sequencing a current goes through nanopores and as nucleotides travel through the nanopores, the resulting change in the current is measured and recorded. The data generated then undergoes basecalling, which is the process in which the changes in current are translated into nucleotide sequences. This analysis can occur simultaneously during sequencing or after sequencing completion [9]. Currently, ONT offers the MinION sequencing instrument and MinION flow cells which contain 4x215 nanopore channels, each containing one nanopore. Additionally, ONT offers the Flongle adapter and Flongle flow cells which contain 126 nanopore channels and are used in addition to the MinION [10]. Larger throughput instruments such as the GridION, which supports sequencing with five MinION flow cells, and the PromethION, which supports the PromethION flow cell containing 3,000 nanopore channels are also available [11]. Although

the initial instrument cost is relatively low for the MinION and Flongle adapter and their associate flow cells, library preparation costs remain high [12]. Despite library preparation costs, the lower initial investment allows for sequencing to be available to a wider range of laboratories. The MinION sequencer is a palm-sized device and as such is a portable instrument allowing sequencing to be performed outside the laboratory, opening up the potential for bed-side diagnostics. The major disadvantage of ONT sequencing is low read accuracy, which has been reported to be 90% with the R 9.4 flow cells [13]. Newer flow cells released by ONT have reported higher accuracies of 87-98% [14, 15]. Read accuracy of 99% has also been reported, however, this was achieved after the addition of bioinformatic polishing which is highly computation and time intensive [16, 17].

7.1.1 ONT Flongle

For diagnostic purposes, high throughput data is not required to achieve accurate identification. Previous studies have determined that nanopore sequencing with the MinION identified bacterial pathogens within 30 minutes of sequencing and an overall turnaround time of 6-8 hours from sample collection to diagnostic report [18-20]. As such, the smaller Flongle flow cells from ONT are suitable for diagnostic applications with smaller throughput requirements. The lower cost and pore number of the flow cells are appropriate for sequencing a single sample and reduce the need for multiplexing of samples to justify the cost of higher throughput sequencing. Additionally, in diagnostics, rapid identification is essential and so waiting for an adequate number of patient samples to be multiplexed and sequenced together is not desirable for optimal patient outcomes.

Flongle flow cells have been used in previous studies to achieve species identification and generate further genetic information from bacterial and viral pathogens. Common bacterial pathogens such as *Escherichia coli* and *Klebsiella pneumoniae* have been identified in spiked

samples, such as milk, minced beef, and blood using Flongle flow cells [21-23]. Accurate detection was achieved within three hours across all studies and further information such as antimicrobial resistance genes and characteristic virulence genes was identified. Complete genome sequences of *Pseudomonas coronafaciens* pv. Oryzae 1_6, *Lactobacillus iners*, and 20 actinomycete strains were generated by nanopore sequencing with Flongle flow cells, either alone or in conjunction with short-read Illumina sequencers [24-26]. Most notably, Flongle flow cells have been evaluated on their use for the identification and characterisation of the SARS-CoV-2 virus [27, 28]. Variants in the partial *RdRP* gene and immunodominant part of the *S-gene* were successfully amplified and sequenced on Flongle flow cells which then had the potential to inform on increased infectivity and potential immune evasion of the SARS-CoV-2 virus [27, 28]. Rapid identification and genotyping of enterovirus species were also achieved with nanopore sequencing with Flongle flow cells, and the identification of viral plant pathogens [29, 30].

The use of ONT sequencing using a metagenomic approach in fungal identification is limited and in studies where this has been explored, MinION flow cells were largely used [31, 32]. A study investigating the gastrointestinal mycobiome of indoor and outdoor felines exclusively used Flongle flow cells [33]. DNA extracted from ten feline stool samples underwent metabarcoding with the ITS1 region. Diversity was identified at the genus level and was found to have differences between indoor and outdoor cats, with an outdoor environment suggesting greater mycobiome diversity [33]. The application of Flongle flow cells to metagenomic identification of pathogenic fungi in patient samples has currently not yet been explored.

7.2 Methods

7.2.1 Sample acquisition and routine diagnosis

Three samples were collected from patients with suspected fungal infections and routine diagnosis was performed to identify the causative pathogens (P1, P2 and P3) (Table 1). P1 was a sample of body fluid drained by a lump on the ankle of a 60-year-old female that had undergone renal transplantation. Potential pathogens were detected through culturing of body fluid. P2 was a sample of lung tissue from a 28-year-old male and underwent culturing for fungal growth and *Aspergillus* specific PCR. P3 was a sample of CSF from a 21-year-old female on which panfungal PCR was used for diagnosis.

7.2.2 DNA Extraction for metagenomic sequencing

DNA was extracted with the Invitrogen by Thermo Fischer Scientific PureLink™ Genomic DNA Mini Kit (Cat. No. K1820-01) following a modified version of the manufacturer's instructions.

DNA extraction was performed on 200 µL of body fluid from P1, all lung tissue available from P2, and 200 µL of CSF from P3. Patient samples were initially frozen with liquid nitrogen and crushed with a pestle to a fine powder to break fungal cell walls. Samples were then divided to undergo suitable lysate preparation steps as described in the manufacturer's instructions.

P1 and P3 underwent lysate preparation under instructions for mammalian cells. Patient samples were thawed at room temperature and 20 µL Proteinase K and 20 µL RNase A were added to the sample. The solution was mixed well by vortexing and incubated at room temperature for 2 minutes. 200 µL PureLink Genomic Lysis/Binding Buffer was added to the solution and mixed to obtain a homogenous solution by vortexing. The solution was then

incubated at 55°C for 10 minutes. 200 µL 100% ethanol was added and mixed by vortexing to obtain a homogenous solution and then proceeded to the remainder of the protocol.

P2 underwent lysate preparation under instructions for mammalian tissue. 180 µL of PureLink Genomic Digestion Buffer and 20 µL Proteinase K were added to the sample tube. The sample was then incubated at 55°C for 1.5 hours, vortexing every 30 minutes. The solution was centrifuged for 3 minutes at maximum speed and the supernatant was transferred to a sterile microfuge tube. 20 µL RNase was added to the solution, mixed well by vortexing, and incubated at room temperature for 2 minutes. 200 µL PureLink Genomic Lysis/Binding Buffer was added to the solution and mixed well by vortexing. 200 µL 100% ethanol was added, mixed by vortexing, and then proceeded to the remainder of the protocol.

P1, P2, and P3 underwent the same process for the remainder of DNA extraction. After lysate preparation, the samples were pipetted into PureLink Spin Columns and centrifuged at 10,000 *xg* at room temperature for 1 minute. The supernatant and collection tube were then discarded, and the spin column was then placed into a new PureLink Collection Tube. 500 µL Wash Buffer 1 was added to the column and centrifuged at 10,000 *xg* at room temperature for 1 minute. The collection tube was discarded, and the spin column was placed into a new PureLink Collection Tube. 500 µL Wash Buffer 2 was added to the column and centrifuged at maximum speed at room temperature for 3 minutes. The collection tube was discarded, and the spin column was placed into a sterile 1.5 mL microfuge tube. 200 µL nuclease-free water was added to the column and incubated at room temperature for 1 minute. The column was then centrifuged at maximum speed at room temperature for 1 minute. Elution was repeated with 200 µL nuclease-free water added to the column and incubated at room temperature for 1 minute. The column was then centrifuged at maximum speed at room temperature for 1 minute.

The volume of extracted DNA was reduced to 50 µL in a concentrator at 45°C. The resulting DNA concentration was measured with the DeNovix dsDNA Broad Range Assay on the DeNovix DS-11 FX+ and resulting DNA purity was measured through A260/A230 and A260/A280 values measured on the DeNovix DS-11+. DNA was stored at 4°C for further use.

7.2.3 Library Preparation

DNA sequencing libraries were prepared using the Ligation Sequencing Kit SQK-LSK109 from ONT following the manufacturer's instructions.

DNA underwent repair and end-preparation by the addition of 1 µL DNA CS, 3.5uL NEBNext FFPE Repair buffer, 2 µL NEBNext FFPE DNA Repair Mix, 3.5 µL Ultra II End-prep reaction buffer, and 3 µL Ultra II End-prep enzyme mix to 47 µL of extracted sample DNA. For library preparation of P2 and P3, 1 µL DNA CS was not added to the solution and 1 µL nuclease-free water was added instead. The solution was then gently mixed by flicking, spun down, and incubated at 20°C for 5 minutes and 65°C for 5 minutes. 60 µL of resuspended AMPure XP beads was added to the solution, gently mixed, and incubated whilst rotating at room temperature for 5 minutes. The sample was then spun down and beads were pelleted on a magnet. The supernatant was removed, and the pellet was washed with 200 µL of 70% ethanol twice. The pellet was then spun down and pelleted on a magnet. Any remaining ethanol was removed by pipetting and the pellet was dried at room temperature for 30 seconds. The pellet was resuspended in 61 µL nuclease-free water and incubated at room temperature for 2 minutes. The beads were pelleted on a magnet until the eluate was clear and colourless. The eluate was then transferred to a clean 1.5 mL Eppendorf LoBind

microcentrifuge tube. 1 µL was taken to measure DNA concentration with the DeNovix dsDNA Broad Range Assay on the DeNovix DS-11 FX+.

The DNA library then underwent adapter ligation by the addition of 25 µL Ligation Buffer, 10 µL NEBNext Quick T4 DNA Ligase and 5 µL Adapter Mix to the 60 µL of the DNA library. The solution was then mixed gently and incubated at room temperature for 10 minutes. 40 µL of resuspended AMPure XP beads were added to the solution, mixed gently, and incubated whilst rotating at room temperature for 5 minutes. Beads were pelleted on a magnet and the supernatant was removed. 250 µL Short Fragment Buffer was added to the solution and gently flicked until beads were resuspended. The beads were then pelleted on a magnet and the supernatant was discarded. The beads were then resuspended in 250 µL of Short Fragment Buffer by gentle flicking and pelleted on a magnet. The supernatant was discarded, and the pellet was resuspended in 15 µL Elution Buffer. The solution was incubated at room temperature for 10 minutes. Beads were then pelleted on a magnet until the eluate was clear and colourless. Eluate was transferred to a clean 1.5 mL Eppendorf LoBind microcentrifuge tube and beads were discarded. 1 µL was taken to measure DNA concentration with the DeNovix dsDNA Broad Range Assay on the DeNovix DS-11 FX+.

7.2.4 DNA Sequencing

Sequencing of each patient sample was performed on Flongle flow cells (FLO-FLG001) (R9.4.1) connected to a Flongle adapter within a MinION Mk1B. Flow cells were primed with a priming mix composed of 117 µL Flush Buffer and 3 µL Flush Tether. A total volume of 30 µL sequencing mix, composed of 15 µL Sequencing Buffer, 10 µL Loading Beads and 5 µL prepared DNA library, was prepared and loaded into the Flongle flow cell using a modified loading protocol by G. Wiley [34]. Sequencing was then executed through MinKNOW and sequencing was run for 24 hours or until no sequencing pores remained.

Raw fast6 files were basecalled through Albacore 2.3.1 software and reads were also filtered to exclude reads with Phred Quality (Q) score <7, as per the manufacturer's recommendations.

7.2.5 Data Analysis

The same methodology for data analysis was used as those described in Chapter 5 under the heading "Data analysis" using the WIMP analysis tool.

7.3 Results

7.3.1 Routine Diagnosis of Patient Samples

Routine diagnosis of all three patient samples revealed the presence of fungal infections for each patient. Culturing of body fluid from P1 revealed the growth of the fungal species *Medicopsis romeroi.* Fungal growth from P2 tissue was identified as *Aspergillus flavus* and the patient sample was positive for *Aspergillus* specific PCR. Panfungal PCR of P3 revealed the sample was positive for *Cryptococcus gattii* DNA.

7.3.2 Sample DNA Concentration and Purity

The yield of DNA from the extraction of patient samples led to varying results with the highest yield of 1466 ng from P1, 463 ng from P2, and the lowest 146 ng from P3. To determine the purity of samples, A260/A230 and A260/A280 values were measured with values 1.80 – 2.20 indicating high sample purity. The A260/A230 values for all three patient samples were <1.80 (Table 1). The A260/A280 values for P1 and P2 were 1.85 and 1.82 respectively and the value for P3 was 1.52.

Table 1: Quantification of extracted DNA purity from patient samples, using A260/A230 and A260/A280 values measured by spectrophotometry.

Sample	A260/A230	A260/A280
P1	1.42	1.85
P2	0.48	1.82
P3	0.55	1.52

The total amount of DNA extracted from the patient samples was used in library preparation. After library preparation, P1 retained 380 ng of DNA and 136 ng was loaded for sequencing. The final amount of DNA after library preparation for P2 was 120 ng and 43 ng of DNA was loaded for sequencing. For P3, 60 ng of DNA was remaining after library preparation and 22 ng of DNA.

7.3.3 Metagenomic Sequencing

Albacore software was used to basecall the raw sequencing data to nucleotide sequences. Q scores were used to determine the quality of the sequencing reads and Albacore divided reads into pass (Q score > 7) and fail (Q score ≤ 7). Reads that belonged to the pass group underwent further analysis.

Sequencing of the P1 sample was initially done with 58 nanopore channels available for sequencing after library loading. The P1 sample yielded a total of 443,348 reads and 341.0 Mbps, with 81.8% of reads passing the Q score filter, amounting to 362,866 reads. The average Q score for all reads was 8.98 and there was an average sequencing length of 769 bp.

Sequencing of the P2 sample began with 49 nanopore channels available for sequencing, however after 11 hours of sequencing there was a sharp decrease in sequencing pores and no pores remained viable after 15.5 hours (Figure 1). The P2 sample yielded a total of 2,975 reads and 8.6 Mbps, with 78.9% of reads passing the Q score filter, amounting to 2,348 reads.

The average Q score for all reads was 8.99 and there was an average sequence length of 2,894 bp.

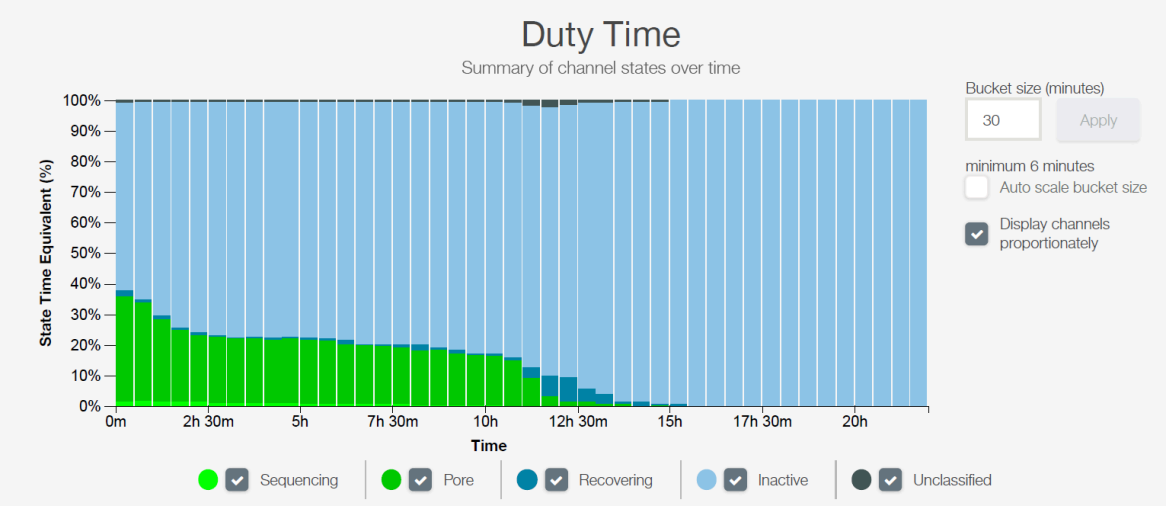


Figure 1: Summary of nanopore channels during sequencing of sample P2. Sequencing of sample P2 began with 49 nanopore channels available for sequencing with steady decrease until 11 hours. Afterward there was a sharp decrease in sequencing pores until all pores were inactive at 15.5 hours.

Sequencing of the P3 sample began with 59 nanopore channels available for sequencing after library loading. The P3 sample yielded a total of 125,420 reads and 311.9 Mbps, with 80.3% of reads passing the Q score filter, amounting to 100,695 reads. The average Q score for all reads was 8.54 and there was an average sequence length of 2,486.

7.3.4 WIMP Classification

WIMP analysis was performed on all quality filtered sequencing data, and taxon assignment was based on the highest WIMP score for each sequencing read. WIMP classified 91-97% of reads for all samples. Importantly, no samples had reads assigned to the taxon identified as the causative pathogens by routine diagnosis. Reads assigned to *Homo sapiens* represented

>99% of assigned reads for all samples with 99.4%, 99.8%, and 99.9% for samples P1, P2 and P3 respectively.

WIMP analysis of sample P1 reads assigned reads to 130 different taxa (Figure 2A). *Escherichia* spp. Was the second most abundant taxon with 1,210 reads or 0.4% of the total assigned reads. *Malassezia* spp. Was the third most abundant taxon with 179 assigned reads or 0.05% of total assigned reads. All in all, 120 taxa were represented by ten reads or less.

WIMP analysis of sample P2 assigned reads to five different taxa (Figure 2B). *Fusarium* spp. Was the second most abundant taxon with two reads, accounting for 0.09% of total assigned reads. *Malassezia* spp. was also assigned to one read. All taxa aside from *H. sapiens* were represented by two reads or less.

WIMP analysis of sample P3 assigned reads to 47 different taxa (Figure 2C). *Malassezia* spp. was the second most abundant taxon with 18 reads, accounting for 0.02% of total assigned reads. 45 taxa were represented by ten reads or less and 44 taxa were represented by five reads or less.

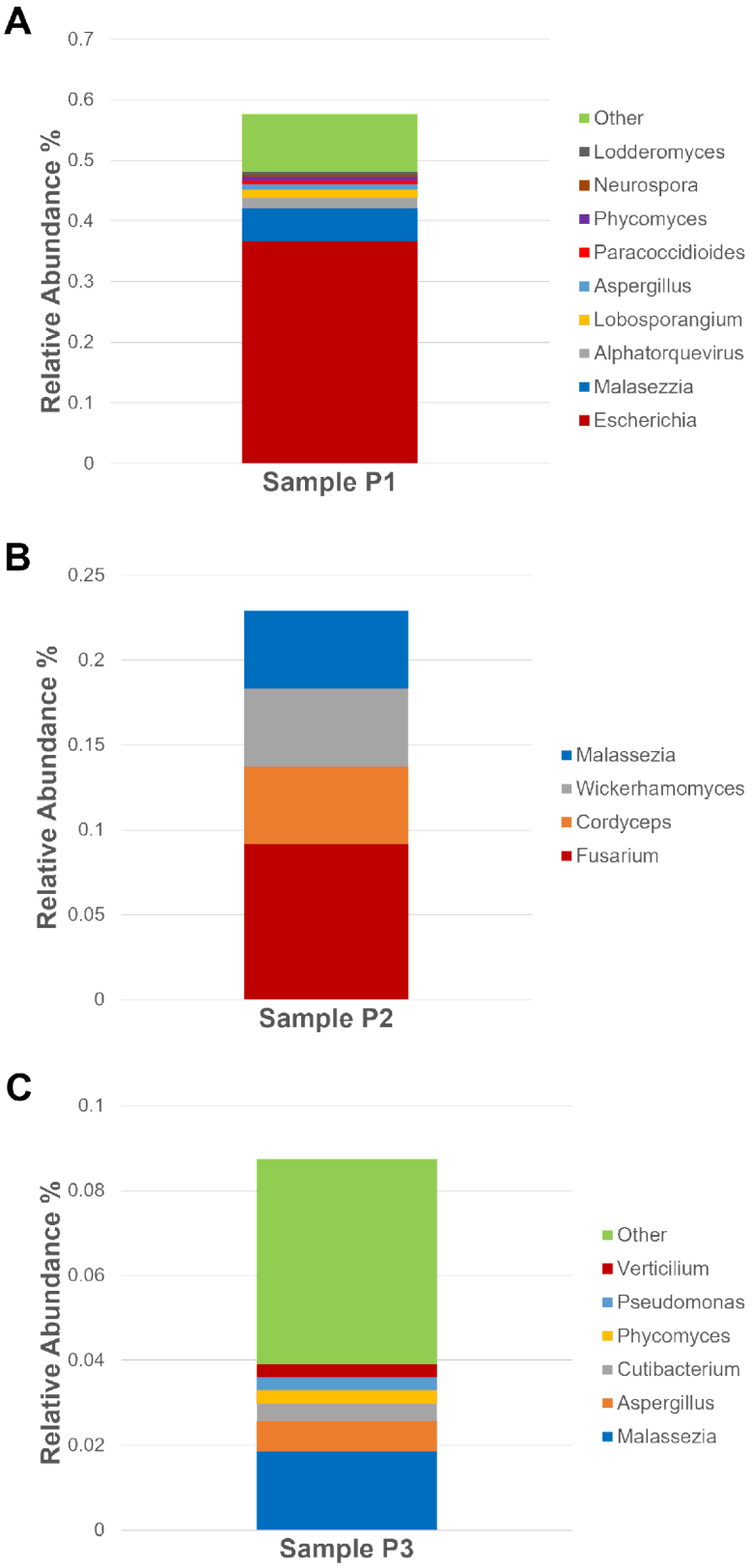


Figure 2: The relative abundance of reads assigned to the most abundance taxon by WIMP classification for nanopore sequencing of patient samples P1(A), P2(B), and P3(C), excluding those assigned to *Homo sapiens*.

7.4 Discussion

The aim of this study was to apply metagenomic sequencing with long-read ONT Flongle flow cells to identify fungal pathogenic species from sterile and non-sterile patient samples. Additionally, the study aimed to determine whether results from metagenomic sequencing with Flongle flow cells were capable of confirming the diagnosis of fungal pathogens obtained through routine diagnostic tests. Fungal pathogens identified in three patient samples using routine diagnostic techniques were unable to be identified in sterile and non-sterile patient samples. This study highlights the developments required to improve the workflow to identify fungal pathogens from patient samples using Flongle flow cells.

Flongle flow cells are the newest sequencing platform released by ONT in 2019 [10]. They have previously been applied to the identification of prominent bacterial species such as, *E. coli* and *K. pneumoniae,* and viral pathogens, SARS-CoV-2 and Enterovirus [21, 22, 27-29]. These studies have demonstrated the potential for the application of nanopore sequencing with Flongle flow cells to the identification of fungal species.

7.4.1 DNA Extraction

The success of nanopore sequencing heavily relies upon the length, quantity, and quality of the DNA being sequenced and as such, DNA extraction is integral to ensure successful sequencing with Flongle flow cells.

The DNA extraction method used in this study was composed of a preliminary step to crush fungal cell walls using liquid nitrogen. The addition of this step was included as fungal cell walls are tough and require manual disruption to allow the following extraction reagents to allow DNA to be extracted from fungal cells. This is especially the case for filamentous fungi and fungal spores which may be present in patient's samples. The remainder of the DNA

extraction method relied on a commercial kit focused on the extraction of human cells. This was chosen to maximise DNA extracted from the sample and the initial manual disruption would allow for fungal DNA to be extracted alongside human DNA. An extraction method specifically designed to extract DNA from fungal cells may be an important comparison to determine which commercial kit would more effectively extract fungal DNA from patient samples.

DNA extraction of patient samples in this study resulted in varying yields from 1466-146 ng. This discrepancy is likely due to the use of different types of patient samples undergoing similar DNA extraction methods. Samples P1 and P3 were body fluid and CSF respectively, and although these samples are vastly different, the same DNA extraction method was used for both. As a result, sample P1 had a yield of DNA 10-fold the yield of DNA from sample P3 indicating this method is more suited toward body fluid than CSF. Each sample type would require different extraction methods as each sample site is highly varied with CSF being a sterile site and body fluid being non-sterile. As such, specialised extraction protocols that have been optimised to target specific sample types are required to generate the maximum yield of DNA from patient samples.

The purity of the extracted DNA was determined through the evaluation of the A260/A230 and A260/A280 values generated by spectrophotometry. The measurement of DNA by spectrophotometry is based on the A260 value which is the absorption at 260 nm as at this wavelength nucleic acids naturally absorb light and can be detected [35]. The A260/A230 and A260/A280 values are indicators of purity as they consider the absorption of contaminants at 230 nm and 280 nm. The generally accepted range for the A260/A230 ratio for pure DNA is 1.80 – 2.20 and values outside of this range indicate contamination with phenols, salts, proteins, or polysaccharides [35, 36]. For A260/S280, the accepted range is 1.8 – 2.0, and values outside this range indicate contamination from RNA and proteins [35, 36]. All patient

samples had A260/A230 values < 1.8 and as such indicate contamination likely from compounds remaining from DNA extraction and reagents from the extraction process. Sample P1 had an A260/A280 value of 1.85 and sample P2 had an A260/A280 value of 1.82, falling in the acceptable range for pure DNA, however, sample P3 had a A260/A280 value far above the range for pure DNA indicating large amounts of contamination from RNA and proteins. Further methods to evaluate the purity of DNA are available such as fluorometry and qPCR, however, spectrophotometry ensured minimal sample use which was essential to retain as much DNA yield as possible [35]. To reduce the contamination within DNA for long-read metagenomic sequencing, an additional purification step should be introduced after DNA extraction. Various approaches are available for the purification of DNA such as spin columns available in commercial kits (for example, DNA Clean & Concentrator Kit, Zymo Research), magnetic DNA binding beads (for example, AMPure XP beads, Beckman Coulter), gravity flow columns, phenol-chloroform, dialysis and plug extraction [37].

7.4.2 Metagenomic Sequencing

Metagenomic sequencing runs for samples P1 and P3 produced a higher number of reads and higher base pair yield than sample P2. Although P3 had an initial input of 22 ng of DNA and sample P2 had an initial input of 46 ng of DNA, the total number of reads generated in the sequencing of sample P3 was more than 5-fold that of sample P2 and the base pair yield of sample P3 was more than 36-fold that of sample P2. Additionally, the pores available for sequencing in the Flongle flow cell for sample P2 rapidly declined after 15.5 hours. The low purity of input DNA for sample P2 is most likely responsible for the low yield obtained during the sequencing run and the rapid death of nanopores. The A260/A230 and A260/A280 values for sample P2 fall outside values that indicate high purity and so there was likely a large concentration of contaminants in the extracted DNA remaining after DNA extraction

and library preparation. Contaminants in flow cells interfere with sequencing by causing clogging of pores, making pores unavailable for sequencing, and in high abundance of contaminants will cause pores to become inactive, thereby stopping sequencing. The addition of DNA purification extends the library preparation time however, high purity DNA is essential to utilize nanopore sequencing instruments. The incorporation of DNA extraction methods suitable for the sample type and DNA purification greatly impacts sequencing success by increasing DNA yield and purity and so should be incorporated when sequencing clinical samples using Flongle flow cells.

The mean read length for samples P2 and P3 was over 2000 base pairs long, far exceeding the maximum read length of short-read sequencing technologies. Although sample P1 had the shortest mean read length of 769 bp, this also exceeds the maximum of 600 bp available using short-read sequencing technologies. Additionally, this length spans the full length of the primary and secondary fungal DNA barcodes and other loci that may inform on antifungal resistance or strain typing. As such, sequencing of clinical samples using Flongle flow cells are capable of generating long reads capable of spanning important genetic regions however, different DNA extraction methods may be employed to reduce shearing of DNA fragments, resulting in longer read length.

7.4.3 WIMP Classification and Pathogen Detection

For all three samples, nanopore sequencing using Flongle flow cells was unable to detect the causative pathogen identified by routine diagnostics. Pathogenic fungal species exist at low concentrations in clinical samples making sequencing and identification of causative species difficult without amplification. Reads assigned to *H. sapiens* accounted for >99% of all reads across all three samples. The large amounts of host DNA within the samples made achieving adequate sequencing depth of microbes challenging. This high host: pathogen ratio also

consumes sequencing resources and affects the cost effectiveness of the relatively lost cost nanopore sequencers. The high proportion of host DNA was also observed in previous studies using the higher scale MinION flow cells and this issue is amplified in the Flongle flow cells due to the lower number of nanopore channels. The solution to increase the microbial reads and sequencing depth of potentially pathogenic species when sequencing clinical samples would be to amplify microbial DNA or to deplete host DNA in samples before library preparation and sequencing. Prior amplification of targeted microbial regions, as in metabarcoding studies, would result in more microbial reads, achieving higher coverage of the causative pathogen and increasing the cost-effectiveness more relevant data would be generated. However, prior amplification limits the sequencing information to the targeted loci and introduces potential amplification biases. Alternatively, depletion of host DNA prior does not introduce biases and still allows a metagenomic approach to sequencing, increasing the potential pathogen information generated. Commercial kits are available and laboratory protocols have previously been published, however, the application of such methods to a broad range of clinical specimens has not been explored. To achieve identification of causative pathogens in clinical samples, reduction of host DNA is required and commercial kits or laboratory protocols for human DNA depletion should be employed.

In sample P1, *E. coli* was the second most abundant taxon behind *H. sapiens*, accounting for 0.37% of classified reads. This relatively high abundance of *E. coli* reads was consistent across sequencing results of samples not included in this study. As such, support from ONT provided information that the reagent DNA CS contains Lambda DNA and calibrant strand which had been expressed in *E. coli* resulting in the sequencing of remnant *E. coli.* Library preparation of samples P2 and P3 did not include DNA CS and instead was replaced with nuclease-free water and WIMP analysis did not reveal high any reads assigned to *E. coli.* As

such, further sequencing using the Ligation Sequencing Kit (SQK-LSK109) should omit the addition of DNA CS.

7.4.4 Study Limitations

This study was limited to a small sample size of three patient samples and furthermore, all samples were different types of clinical samples from varying sample sites and had differing causative pathogens identified by routine diagnosis. As such, more specific conclusions such as profiling non-pathogenic microbes that had been sequenced and an abundance of pathogen reads in particular sample types were not possible. This study aimed to apply long-read sequencing using low throughput Flongle flow cells to identify fungal species from clinical samples.

7.5 Conclusions

Metagenomic sequencing with Flongle flow cells using the herein applied DNA extraction methods was not able to identify fungal pathogens from patient samples. Further Improvements to the workflow are needed to enable the identification of fungal species using Flongle flow cells such as the use of DNA extraction methods suited to the sample type and the addition of DNA purification techniques to improve sequencing and limit pore death. Additionally, to improve the host:pathogen ratio, host depletion methods may be implemented. The exclusion of DNA CS during library preparation is essential to prevent the presence of false *E. coli* reads which may produce more reads than the species present in the sample. The implementation of bioinformatic tools specifically designed to identify fungal species and improvements of databases to include more reference fungal sequences would increase the accuracy and confidence of identification of fungal species using Flongle flow cells.

7.6 References

1. Irinyi, L., *et al.*, *DNA barcoding of fungi causing infections in humans and animals.* Fungal Biol., 2016. **120**(2): p. 125-36.
2. Meyer, W., *et al.*, *Database establishment for the secondary fungal DNA barcode translational elongation factor 1α (TEF1α).* Genome, 2019. **62**(3): p. 160-169.
3. Hoang, M.T.V., *et al.*, *Long-reads-based metagenomics in clinical diagnosis with a special focus on fungal infections.* Front. Microbiol., 2021. **12**: p. 708550.
4. Bharti, R. and D.G. Grimm, *Current challenges and best-practice protocols for microbiome analysis.* Brief. Bioinform., 2021. **22**(1): p. 178-193.
5. Garrido-Cardenas, J.A. and F. Manzano-Agugliaro, *The metagenomics worldwide research.* Curr. Genet., 2017. **63**(5): p. 819-829.
6. Couto, N., *et al.*, *Critical steps in clinical shotgun metagenomics for the concomitant detection and typing of microbial pathogens.* Sci. Rep., 2018. **8**(1): p. 13767.
7. Laudadio, I., *et al.*, *Quantitative assessment of shotgun metagenomics and 16S rDNA amplicon sequencing in the study of human gut microbiome.* OMICS, 2018. **22**(4): p. 248-254.
8. Payne, A., *et al.*, *BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files.* Bioinformatics, 2019. **35**(13): p. 2193-2198.
9. Jain, M., *et al.*, *The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community.* Genome Biol., 2016. **17**(1): p. 239.
10. Oxford Nanopore Technologies. *Oxford Nanopore launches Flongle, for rapid, smaller DNA/RNA sequencing tests in any environment*. 2019 28.10.2021]; Available from: https://nanoporetech.com/about-us/news/oxford-nanopore-launches-flongle-rapid-smaller-dnarna-sequencing-tests-any.

11. Oxford Nanopore Technologies. *Products*. 2021 28.10.2021]; Available from: https://nanoporetech.com/products.

12. Tedersoo, L., *et al.*, *Perspectives and benefits of high-throughput long-read Sequencing in microbial ecology.* Appl. Environ. Microbiol., 2021. **87**(17): p. e0062621.

13. Lu, H., F. Giordano, and Z. Ning, *Oxford nanopore MinION sequencing and genome assembly.* Genomics Proteomics Bioinformatics, 2016. **14**(5): p. 265-279.

14. Loit, K., *et al.*, *Relative performance of MinION (Oxford Nanopore Technologies) versus Sequel (Pacific Biosciences) third-generation sequencing instruments in identification of agricultural and forest fungal pathogens.* J. Appl. Environ. Microbiol., 2019. **85**(21): p. 1-20.

15. Logsdon, G.A., M.R. Vollger, and E.E. Eichler, *Long-read human genome sequencing and its applications.* Nat. Rev. Genet., 2020. **21**(10): p. 597-614.

16. Ashikawa, S., *et al.*, *Rapid identification of pathogens from positive blood culture bottles with the MinION nanopore sequencer.* J. Med. Microbiol., 2018. **67**(11): p. 1589-1595.

17. Morrison, G.A., *et al.*, *Nanopore sequencing of the fungal intergenic spacer sequence as a potential rapid diagnostic assay.* J. Clin. Microbiol., 2020. **58**(12): p. 1-16.

18. Sakai, J., *et al.*, *An identification protocol for ESBL-producing Gram-negative bacteria bloodstream infections using a MinION nanopore sequencer.* J. Med. Microbiol., 2019. **68**(8): p. 1219-1226.

19. Taxt, A.M., *et al.*, *Rapid identification of pathogens, antibiotic resistance genes and plasmids in blood cultures by nanopore sequencing.* Sci. Rep., 2020. **10**(1): p. 7622.

20. Schmidt, K., *et al.*, *Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing.* J. Antimicrob. Chemother., 2017. **72**(1): p. 104-114.

21. Buytaers, F.E., *et al.*, *Towards real-time and affordable strain-level metagenomics-based foodborne outbreak investigations using Oxford Nanopore sequencing technologies.* Front. Microbiol., 2021. **12**: p. 1-11.

22. Avershina, E., *et al.*, *Ultrafast and cost-effective pathogen identification and resistance gene detection in a clinical setting using nanopore Flongle sequencing.* Front. Microbiol., 2022. **13**: p. 1-10.

23. Azinheiro, S., *et al.*, *Suitability of the MinION long read sequencer for semi-targeted detection of foodborne pathogens.* Anal. Chim. Acta., 2021. **1184**: p. 1-9.

24. Baltrus, D.A. and M. Clark, *Complete genome sequence of Pseudomonas coronafaciens pv. oryzae 1_6.* Microbiol. Resour. Announc., 2020. **9**(3): p. 1-2.

25. Kwak, W., *et al.*, *Complete genome of Lactobacillus iners KY using Flongle provides insight into the genetic background of optimal adaption to vaginal econiche.* Front. Microbiol., 2020. **11**: p. 1-10.

26. Rajwani, R., *et al.*, *Genome-guided discovery of natural products through multiplexed low-coverage whole-genome sequencing of soil Actinomycetes on Oxford Nanopore Flongle.* mSystems, 2021. **6**(6): p. 1-17.

27. Chan, W.S., *et al.*, *Evaluation on the use of Nanopore sequencing for direct characterization of coronaviruses from respiratory specimens, and a study on emerging missense mutations in partial RdRP gene of SARS-CoV-2.* Virol J, 2020. **17**(1): p. 1-13.

28. Wagner, G.E., *et al.*, *A novel high-throughput nanopore-sequencing-based strategy for rapid and automated S-protein typing of SARS-CoV-2 variants.* Viruses, 2021. **13**(12): p. 1-6.

29. Gradel, C., *et al.*, *Rapid and cost-efficient Enterovirus genotyping from clinical samples using Flongle flow cells.* Genes (Basel), 2019. **10**(9): p. 1-12.

30. Liefting, L.W., D.W. Waite, and J.R. Thompson, *Application of Oxford Nanopore Technology to plant virus detection.* Viruses, 2021. **13**(8): p. 1-17.

31. Irinyi, L., *et al.*, *Long-read sequencing based clinical metagenomics for the detection and confirmation of Pneumocystis jirovecii directly from clinical specimens: A paradigm shift in mycological diagnostics.* Med. Mycol., 2020. **58**(5): p. 650-660.

32. Gu, W., *et al.*, *Rapid pathogen detection by metagenomic next-generation sequencing of infected body fluids.* Nat. Med., 2021. **27**(1): p. 115-124.

33. Tay, D.D., *et al.*, *ITS1 amplicon sequencing of feline gut mycobiome of Malaysian local breeds using Nanopore Flongle.* Arch. Microbiol., 2022. **204**(6): p. 314.

34. Wiley, G. *A very gentle, relatively rapid way to load a Flongle flowcell*. [Community Post] 2019 10/12/2019 [cited 2020 4/2/2020]; Available from: https://community.nanoporetech.com/posts/a-very-gentle-relatively?utm_campaign=K20016%20Digest&utm_source=hs_email&utm_medium=email&_hsenc=p2ANqtz-_RNGARvKUtUdcSg0XxclReE8KtT1kPlQoau6WjgYdWbZo11azLn-C24kaINtNOaA91YYW.

35. Olson, N.D. and J.B. Morrow, *DNA extract characterization process for microbial detection methods development and validation.* BMC Res. Notes, 2012. **5**: p. 668.

36. Sambrook, J. and D.W. Russell, *The condensed protocols from molecular cloning: a laboratory manual*. 2006, Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press.

37. Wang, Y., *et al.*, *Nanopore sequencing technology, bioinformatics and applications.* Nat. Biotechnol., 2021. **39**(11): p. 1348-1365.

Chapter 8
DNA Extraction and Host DNA Depletion Optimization

8.1 Introduction

Fungal bloodstream infections are a major contributing factor to the morbidity and mortality rates of mycoses.

Candida remains a prominent pathogen, especially within health care institutions where it courses e.g., the fourth most common nosocomial infection and second most common nosocomial bloodstream infection in the US [1]. However, recent trends have indicated that hospitalisations are decreasing, particularly in children under 1 year and adults over 65 years [2]. There has also been a shift in the abundance of candidemia causing species with *C. albicans* previously being vastly the most abundant candidemia causing species but is now responsible for less than 50% of candidemia infections. *C. glabrata*, *C. tropicalis*, *C. parapsilosis*, and *C. krusei* have become more prevalent pathogens [3, 4]. Admission to the intensive care unit due to critical illness is a well-described risk factor for candidemia [5]. Other risk factors predispose patients to infection with specific *Candida* species, such as stem cells and whole organ transplant patients most commonly being infected with *C. glabrata* [6].

8.1.1 Fungal Blood Infection Diagnostics

Rapid and accurate diagnosis is essential to improve patient outcomes [7]. Especially amongst *Candida* species, *C. glabrata* and *C. krusei* have high a innate resistance to the common antifungals group fluconazole and so correct identification to enable correct treatment is important [4]. This is difficult to achieve in bloodborne fungal disease and as such a majority of diagnostic techniques rely on growing the pathogen to a level where it can then be detected.

Blood culturing is commonly used within routine diagnostics to identify the causative agents of fungal blood infections such as candidemia. This involves the collection of blood samples

from patients which are then incubated under conditions to encourage microbial growth. The samples are then tested to detect the presence of microbial cells. There are several commercial blood culturing systems available that have been optimised to incubate blood samples alongside nutrient broth to improve microbial growth speed and include a monitoring system so samples can be consistently monitored and alerted when adequate microbial growth is achieved [8]. After positive culturing is achieved, samples then undergo further diagnostic testing to identify the pathogen. These are widely used in routine diagnostic laboratories and some prominent blood culturing systems are the BacT/Alert Virtuo, BD Bactec Plus ad the VersaTREK [8, 9].

Blood culturing first identifies *Candida* cells at a median concentration of 1 CFU/mL. However, 26-65% of positive *Candida* blood samples had concentrations <1 CFU/mL [10, 11]. Although this limit of detection is comparable to detection with PCR, the sensitivity of blood culturing is estimated to be 50% [12]. *Candida* cells predominantly enter the blood stream by travelling from the gastrointestinal mucosa, which then moves to the liver, which functions as a microbial filter, or by direct inoculation by central nervous catheters, which results in a higher fungal load [11, 12]. Additionally, the rapid removal of *Candida* cells from the bloodstream also factors into the limited sensitivity of blood culturing [13]. The major drawback of blood culturing is the long-time associated with all culturing methods with a 2-3 day median time to a positive result, which may be up to 8 days [10, 14, 15]. Different *Candida* species also require longer time to generate a positive result with *C. glabrata* and *C. parapsilosis* requiring longer than *C. albicans* [10, 12]. After culturing, samples undergo additional diagnostic testing such as those discussed in Chapter 1.

8.1.2 Metagenomic Sequencing and Host DNA Depletion

Metagenomic sequencing involves direct sequencing from samples and is a technique that overcomes the limitations of culturing. It has the potential to revolutionise molecular diagnostics due to the bountiful pathogen genetic information that is able to be generated and analysed. This information could then be used to identify the causative pathogen, antimicrobial markers to inform treatment, and epidemiological markers that may be used in further studies. This approach in conjunction with long read sequencers, such as those available by ONT, has further potential to provide information on microbe abundance, as there is no prior amplification step to alter the microbial community composition. As such, the combination of long-read sequencing and a metagenomic sequencing approach presents high potential for fungal diagnosis.

Previous applications of metagenomic sequencing of pathogens from clinical samples were greatly complicated by the overwhelming presence of host DNA. As such, a need to reduce the ratio of the pathogen to host DNA is required to successfully implement metagenomic sequencing in diagnosis.

An approach to improving the ratio is the amplification of fungal DNA within patient samples. This is commonly done through metabarcoding in which a region of DNA is selected, amplified, and sequenced. Although this results in a higher number of microbial reads, this approach eliminates the undiscriminating nature of metagenomics and may lead to false negative results. Metabarcoding also does not take advantage of long-read sequencers as amplified regions are generally far shorter than the ultra-long reads that long-read sequencers are capable of [16, 17]. Through the selection of identifying genomic regions, additional information, such as anti-fungal markers and epidemiological data, is lost. An alternative method is adaptive sequencing with ONT sequencers, in which unwanted sequences are selected, identified, and rejected from sequencing pores before sequencing is completed. This

method has been applied to the characterisation of bacterial communities within clinical samples resulting in increased bacterial sequencing depth whilst community composition remained consistent [18, 19]. Adaptive sequencing has yet to be explored to exclude human DNA from fungal DNA, although this is a promising route for host DNA depletion.

The opposing approach to improve the pathogen to host DNA ratio is to reduce the amount of host DNA or reads present. This is done before sequencing by the elimination of human cells and DNA from the sample before extraction of microbial DNA. There are several commercial kits and laboratory protocols available that are more often designed to isolate bacterial DNA and the application of these methods to fungal pathogens needs to be explored [20-22].

This part of the study aimed to identify and optimise a host depletion method that can be applied to fungal identification using nanopore sequencing to improve the limit of detection of *Candida albicans* spiked blood samples.

8.2 Methods

8.2.1 Comparison of Host DNA Depletion Methods

Four human DNA depletion methods were tested on *Candida albicans* spiked blood samples (Table 1).

8.2.1.1 Blood Spiking

The ATTC 18804 type culture of *Candida albicans* was selected for this study. The selected strain was streaked onto Sabouraud agar and incubated at 27^0C for 24 hours. Sterile blood was collected in 10mL K_2EDTA vacutainer tubes from healthy volunteers. Primary *C. albicans* suspensions were prepared in phosphate buffered saline (PBS) and colony forming units (CFU) were counted in a haemocytometer. These were then diluted to form secondary

suspensions to produce a final concentration of 10^8 CFU in 210µL. Spiked blood samples were made by combining 200µL of sterile blood with 10µL of the secondary *C. albicans* suspensions. Dilutions were validated by CFU count on Sabouraud agar.

8.2.1.2 Host Depletion Methods

Spiked blood samples with a final concentration of 10^8 CFU were then processed through different host depletion methods.

8.2.1.2.1 Untreated

Spiked blood samples that were not pre-treated with any human DNA depletion methods were included in this study as a negative control. After blood spiking, the samples proceeded to manual fungal cell wall disruption.

8.2.1.2.2 Saponin Depletion

Host DNA was depleted using a modified protocol based on Charalampous *et al.* [21]. The spiked blood samples were centrifuged at 8,000 *xg* for 5 minutes. The upper layer was removed and saponin was added to a final concentration of approximately 2.5%. This was performed by the addition of 100 µL of 5% saponin. The solution was mixed well and incubated at room temperature for 30 minutes. 350 µL of nuclease free water was added to the solution and then incubated for 30 seconds. 12 µL of 5M NaCl was added and mixed briefly by vortex. 100 µL HL-SAN Buffer (5.5M NaCl, 100mM $MgCl_2$) and 10 µL HL-SAN DNase (25,000 units) were added to the solution and incubated at 37°C for 30 minutes, shaking at 800 rpm.

8.2.1.2.3 Buffer RLT Depletion

The depletion of host human DNA using Buffer RLT was based on the study by Rodriguez *et al.* [23]. Spiked blood samples were centrifuged for 10 minutes at 20,000 *xg* and the supernatant was removed. The pellet was resuspended in a 1mL aliquot of Buffer RLT, supplemented with 1% β-mercaptoethanol. The solution was incubated for 1 minute at room temperature and then centrifuged for 8 minutes at 20,000 *xg*. The supernatant was pipetted and removed. The remaining pellet was resuspended in 100 µL of nuclease free water.

2.3.1.2.4 Plasma Depletion

The study by Gu *et al.* reduced the proportion of human DNA sequenced from blood samples by way of isolating plasma [24]. Spiked blood samples were centrifuged for 10 minutes at 6,000 *xg* within 6 hours of collection. The upper plasma layer was pipetted into a new microfuge tube and was then centrifuged for an additional 10 minutes at 16,000 *xg*.

8.2.1.2.5 Molzym Depletion

Removal of human DNA was performed according to the host depletion section of Protocol 1 of the Molzym Ultra-Deep Microbiome Prep10 commercial kit and as briefly described here. *Candida albicans* spiked blood was made up to 10 mL with buffer SU, added to 4 mL buffer *CM* and incubated at room temperature for 5 minutes. 4 mL Buffer *DB1* and 10 µL *MolDNase B* were added and the solution as incubated at room temperature for 15 minutes. The mixture was then centrifuged at 9,000 *xg* and the solution was removed from the top, leaving behind approximately 200 µL. The remaining solution was transferred to a 1.5 mL microfuge and centrifuged at 12,000 *xg* for 5 minutes. The supernatant was pipetted off and resuspended in 80 µL buffer *RL*.

8.2.1.3 Manual Fungal Cell Wall Disruption

Fungal cell walls were manually crushed to release pathogen DNA. Plastic pestles were placed into microfuge tubes and submerged in liquid nitrogen. The pestles were then rotated until the samples were crushed to a power and resubmerged in liquid nitrogen if required. Samples were thawed at room temperature and plastic pestles were discarded.

8.2.1.4 DNA Extraction

DNA was extracted using the QIAGEN DNA Mini Blood Kit as briefly described below. 20µL QIAGEN Protease and 200 µL Buffer AL were added to the sample and mixed by vortex for 15 seconds. The samples were then incubated at 56°C for 10 minutes. 200 µL 96-100% Ethanol was added and mixed by vortex. The solutions were then spun through QIAGEN QIAamp Mini spin columns at 6,000 *xg* for 1 minute. 500 µL Buffer AW1 was added and spun through the column at 6,000 *xg* for 1 minute. 500 µL Buffer AW2 was added and spun through the spin column at 16,000 *xg* for 3 minutes. The column was then spun for an additional 1 minute at maximum speed. 200 µL nuclease free water was applied to the column, incubated at room temperature for 1 minute and centrifuged at 6,000 *xg* for 1 minute to elute DNA. This elution step was repeated, and the resulting DNA pooled together and concentrated to 50 µL. The DNA was stored at 4°C until used.

8.2.1.5 DNA Concentration and Purity

Host depletion methods were compared by the quantity and quality of DNA extracted. The concentration of DNA was measured by fluorescence with the DeNovix Broad Spectrum Kit on the DeNovix DS-11 FX+ and by spectrophotometry on the DeNovix DS-11+. The purity of the extracted DNA was quantified by the A260/A230 and A260/A280 values measured on the DeNovix DS-11+ (Table 2).

8.2.2 Limit of Detection of untreated vs saponin treated *C. albicans* spiked blood

The limit of detection of *C. albicans* DNA by nanopore sequencing on Flongle flow cells was compared for untreated and saponin treated blood samples.

8.2.2.1 Sample preparation

Spiked *Candida albicans* blood samples were prepared in serial ten-fold dilution from 10^4CFU to 10^9CFU as described in Section 8.1.1.1. Saponin treated samples were processed as described in Section 8.1.1.2.2. Both untreated and saponin treated samples had DNA extracted as described in Section 8.1.1.3 and Section 8.1.1.4. DNA concentration and purity were measured as described in Section 8.1.1.5.

8.2.2.2 Nanopore Sequencing

Samples were prepared for sequencing by library preparation using the Ligation Sequencing Kit (SQK-LSK109) in accordance with the manufacturer's instructions. This is described in Chapter 7 Section 7.2.4. Each sample was sequenced in an individual Flongle flow cell for 24 hours or until the flow cell ceased sequencing.

8.2.2.3 Data Analysis

Sequencing data underwent analysis using the WIMP analysis tool as previously described in Chapter 7 under the heading "Data analysis".

8.2.3 Optimisation of Saponin Human DNA Depletion

The modified saponin host DNA depletion method from Charalampous *et al.* underwent optimization to increase the percentage of sequencing reads assigned to *Candida*, as such improving the human DNA depletion.

8.2.3.1 Blood Spiking

The *C. albicans* type strain (ATTC 18804) was streaked onto Sabouraud agar and incubated at 27^0C for 24 hours. Primary *C. albicans* suspensions were prepared PBS and CFU were counted in a haemocytometer. The primary suspensions were then further diluted to produce secondary suspensions to achieve final concentrations of 10^4 CFU to 10^7 CFU in 5.1 mL. Sterile blood was collected in 10 mL K_2EDTA vacutainer tubes from healthy volunteers. 5 mL of sterile blood was then combined with 100 µL of the appropriate secondary suspension to produce a spiked blood sample with a final *C. albicans* concentration of 10^4 CFU – 10^7 CFU in 5.1 mL. Dilutions were validated by CFU count on Sabouraud agar.

8.2.3.2 Saponin Host Depletion Method

Three modifications of the saponin human DNA method from Charalampous *et al* were tested on *C. albicans* spiked blood samples.

8.2.3.2.1 Untreated

Spiked blood samples that were not pre-treated with any human DNA depletion methods were included in this study as a negative control. After blood spiking, the samples proceeded to manual fungal cell wall disruption.

8.2.3.2.2 Saponin Depletion Method 1

The spiked blood samples were centrifuged at 3,000 *xg* for 10 minutes. The upper layer was removed and saponin was added to a final concentration of approximately 2.5% by the addition of 5 mL 5% saponin. The solution was mixed well by vortexing and incubated for 30 minutes at room temperature. 350 µL of nuclease free water was added to the solution and incubated for 30 seconds at room temperature. 12 µL of 5M NaCl was added and mixed. 100 µL HL-SAN Buffer (5.5M NaCl, 100mM $MgCl_2$) and 10 µL HL-SAN DNase (25,000 units) were added and incubated at 37°C for 30 minutes, shaking at 800 rpm.

8.2.3.2.3 Saponin Depletion Method 2

The second saponin depletion method was largely performed as described in section 8.2.3.2.2 however, after the addition of saponin to a final concentration of 2.5%, the solution was incubated for 45 minutes at room temperature. The volume of nuclease free water added was increased to 750 µL and the volume of 5M NaCl was increased to 50 µL. The final incubation was also prolonged to 40 minutes at 37°C, shaking at 800rpm.

8.2.3.2.4 Saponin Depletion Method 3

Spiked blood samples were centrifuged at 3,000 *xg* for 10 minutes. The upper layer was discarded by pipetting and saponin was added to produce a final concentration of approximately 5% through the addition of 5 mL 10% saponin solution. The solution was then mixed well and incubated at room temperature for 45 minutes. 1 mL of nuclease free water was then added to the solution and incubated at room temperature for 1 minute. 75 µL of 5M NaCl was added to the solution. 200 µL HL-SAN buffer (5.5M NaCl, 100mM $MgCl_2$) and 20 µL HL-SAN DNase (50,000 units) were added and mixed thoroughly. The solution was then incubated at 37°C for 1 hour, shaking at 800 rpm.

8.2.3.3 Manual Fungal Cell Wall Disruption

Fungal cell walls were manually crushed using liquid nitrogen. The spiked blood was poured into a mortar and liquid nitrogen was added gradually. The solutions were then crushed using a pestle to form a powder and poured into 50 mL tubes whilst still a powder. The blood solution was then thawed at room temperature.

8.2.3.4 DNA Extraction

DNA was extracted using the QIAGEN DNA Maxi Blood Kit as briefly described below.

Spiked blood samples that were untreated or underwent saponin depletion method 1 or saponin depletion method 2 were processed with instructions for a starting volume of 3-5 mL. Spiked blood samples that underwent saponin depletion method 3 had 3 mL PBS added and were processed with instructions for a starting volume of 5-10 mL which followed the same protocol with increased volumes of solutions added as described below.

6 mL of Buffer AL was added to the spiked blood samples and mixed by inversion, followed by vortexing for 1 minute. For saponin depletion method 3, the volume of Buffer AL added was increased to 12 mL. The solution was then incubated at 70°C for 10 minutes. 5 mL of 96-100% ethanol was then added and mixed by inversion and vortexing. For saponin depletion method 3, 10 mL of ethanol was added instead. The solution was then spun through QIAamp Maxi columns by centrifuging at 1,850 *xg* for 3 minutes and the filtrate was discarded. 5 mL Buffer AW1 was added to the QIAamp Maxi column and centrifuged at 4,500 *xg* for 15 minutes and the filtrate was discarded. 5 mL Buffer AW2 was added to the QIAamp Maxi column and centrifuged at 4,500 *xg* for 15 minutes and the filtrate was discarded. The QIAamp Maxi column was placed in a new 50 mL centrifuge tube and 600 µL nuclease free water was added to the QIAamp Maxi column and after incubation at room temperature for 5 minutes, the column was centrifuged at 4,500 *xg* for 2 minutes. 600 µL

nuclease free water was again added to the QIAamp Maxi column and after incubation at room temperature for 5 minutes, the column was centrifuged at 4,500 *xg* for 5 minutes. For saponin depletion method 3, the volume of nuclease free water was increased to 1 mL for both elution steps.

Extracted DNA was then concentrated to 50 µL and stored at 4°C for further use.

8.2.3.5 DNA Purification

Samples that were treated using saponin depletion methods 2 and 3 underwent further DNA purification steps following DNA extraction using the Zymo Research Genomic DNA Clean and Concentrator. Saponin depletion method 3 samples also underwent additional DNA purification using AMPure beads.

8.2.3.5.1 Zymo Research Genomic DNA Clean and Concentrator

175 µL ChIP DNA Binding Buffer was added to extracted DNA samples and mixed by vortex. The solution was then transferred to a Zymo-Spin™ IC-XL Column in a collection tube, centrifuged at maximum speed for 30 seconds and the flow through was discarded. 200 µL DNA Wash Buffer was added to the column, the column was then centrifuged at maximum speed for 1 minute. This last step was then repeated. 50 µL nuclease free water was then added to the column and incubated at room temperature for 1 minute. The column was then transferred to a 1.5 mL microfuge tube and centrifuged at maximum speed for 30 seconds. Purified DNA was then stored at 4°C for further use.

8.2.3.5.2 AMPure Bead Cleanup

AMPure beads were resuspended by brief vortexing. 90 µL AMPure beads were added to the DNA sample and mixed to form a homogenous solution by gently flicking. The solution was then incubated at room temperature on a hula mixer for 5 minutes. The AMPure beads were then pelleted on a magnet and the supernatant was removed. Beads were then washed with 70% ethanol twice and dried for 30 seconds. The beads were then resuspended in 50 µL nuclease free water and incubated at room temperature for 10 minutes. Beads were pelleted on a magnet and eluate was transferred to a 1.5 mL microfuge tube. Purified DNA was then stored at 4°C for further use.

8.2.3.6 DNA Concentration and Purity

Extracted DNA concentration was then measured by fluorescence, using the DeNovix Broad Spectrum Kit on the DeNovix DS-11 FX+. The purity of samples was also quantified by the A260/A230 and A260/A280 values measured on the DeNovix DS-11+.

8.2.3.7 Nanopore Sequencing and Data Analysis

Library preparation, sequencing and data analysis were performed identically to methods outlined in sections 8.2.2.2 and 8.2.2.3.

8.3 Results

8.3.1 Comparison of host depletion methods

Extracted DNA from the four human DNA depletion methods, saponin depletion, Buffer RLT depletion, Plasma depletion, and Molzym depletion, were compared to untreated extracted DNA. Two samples for each host depletion treatment were processed and their concentration and purity were measured. The concentration of DNA by fluorescence was highest in the

untreated condition with 2.84-2.99 ng/µL, followed by the saponin depletion with 1.61-1.51 ng/µL (Table 1). All other fluorescence measurements indicated that the concentration was too low to be detected with the DeNovix Broad Spectrum Kit on the DeNovix DS-11 FX+. All A260/A230 measurements were <1.8 which is the standard indicating high purity. Additionally, the A260/A280 values for all host depletion treatments and the untreated condition, except for the Molzym depletion were >1.8. The Buffer RLT and plasma depletion conditions had large ranges for the A260/A280 values with ranges of 3.88 and 4.47 respectively. The untreated and saponin depletion methods had smaller A260/A280 ranges with 0.03 and 0.02 respectively.

Table 1: The concentration and purity of DNA extracted from 10^8 CFU *C. albicans* spiked blood without human DNA depletion and with saponin depletion, buffer RLT depletion, plasma depletion, and Molzym depletion.

Host Depletion Treatment	Repeat	Concentration by Fluorescence (ng/µL)	A260/A230	A260/A280
Untreated	1	2.99	0.58	2.03
	2	2.84	0.67	2.06
Saponin Depletion	1	1.61	0.36	1.82
	2	1.59	0.36	1.84
Buffer RLT Depletion	1	Too Low	0.31	3.91
	2	Too Low	0.56	7.79
Plasma Depletion	1	Too Low	0.35	1.77
	2	Too Low	0.27	6.24
Molzym Depletion	1	Too Low	0.21	0.70
	2	Too Low	0.16	1.32

8.3.2 Limit of Detection

The limit of detection of *C. albicans* in spiked blood samples was investigated for untreated samples and saponin depletion samples. Detection was distinguished by reads assigned to

Candida spp. , being the most abundant within a sample, or the second most abundant, after reads assigned to *H. sapiens*.

8.2.3.1 Untreated

The DNA yield before library preparation and quantity of DNA loaded for sequencing for concentrations 10^9 CFU to 10^4 CFU that did not undergo any human DNA depletion methods are contained in Table 2.

The sequencing run of the 10^9 CFU spiked blood sample produced 10,325 reads that passed the Albacore quality score and 9965 reads were classified by WIMP. Reads classified to *H. sapiens* were the most abundant, accounting for 5.385 reads, followed by *E. coli* reads numbering 1,881 and *Candida* spp. reads numbering 1,592.

The 10^8 CFU spiked blood sample generated 25,179 reads that passed the Albacore Q score filter and 23,020 reads were classified by WIMP. Reads assigned to *H. sapiens* numbered 19,080 reads and were the most abundant taxon, followed by 2,107 *E. coli* reads and 355 *Candida* spp. reads.

The 10^7 CFU spiked blood sample sequencing run produced 66,238 reads that passed the Albacore Q score filter and 63,483 reads were classified by WIMP. Reads classified as *H. sapiens* had the highest abundance, followed by *E. coli*, numbering 2,082 reads, and *Candida* spp. numbering 148 reads.

Sequencing of 10^6 CFU spiked blood sample generated 69,645 quality reads and of these 66,861 were classified by WIMP. *H. sapiens* was the most abundant taxon with 62,663 reads assigned and *Candida* spp. was the fifth most abundant taxa with 11 reads.

The 10^5 CFU spiked blood sample sequencing produced 72,962 reads further analysed after passing the Albacore Q score filter and 70,100 reads were classified. The most abundant

taxon was *H. sapiens* with 66,357 assigned reads and *Candida* spp. was the fifth most abundant taxon with 8 reads.

Sequencing of 10^4 CFU spiked blood sample produced 77,787 quality reads and of these 73,934 were classified by WIMP. *H. sapiens* was the most abundant taxon with 73,767 reads whilst *Candida* spp. was the eighth most abundant taxon with 4 reads.

The high abundance of *E. coli* reads present in 10^9 CFU,10^8 CFU and 10^7 CFU was likely due to the inclusion of the reagent DNA CS during library preparation as discussed in Chapter 7 Section 7.4.3 and as such, *Candida* spp. Was considered the second most abundant taxa for those sequencing runs. As such, the limit of detection identified for the untreated samples was 10^7 CFU.

Table 2: DNA yield of *C. albicans* spiked blood untreated and treated by saponin depletion at concentrations 10^9 CFU to 10^4 CFU before nanopore sequencing library preparation and after library preparation, to be loaded for sequencing.

Host depletion method	***C. albicans* concentration (CFU)**	**DNA yield before library preparation (ng)**	**DNA loaded for sequencing (ng)**
Untreated	10^9	335.580	26.000
	10^8	535.800	40.040
	10^7	640.281	100.795
	10^6	817.612	109.390
	10^5	732.448	103.760
	10^4	467.650	56.000
Saponin depletion	10^9	169.2	41.15
	10^8	75.67	25.85
	10^7	47.47	32.4
	10^6	36.895	6.9
	10^5	662.7	114.5
	10^4	49.82	40.3

8.2.3.2 Saponin treated

The amount of DNA for concentrations 10^9 CFU to 10^4 CFU of *C. albicans* spiked blood samples, before library preparation and before loading flow cells are described in Table 2.

Sequencing of 10^9 CFU spiked blood sample produced 313,901 reads that passed the Albacore Q score filter, of which 95,234 reads were successfully classified by WIMP. *Candida* spp. was the most abundant taxon with 78,773 reads.

The 10^8 CFU spiked blood sample sequencing generated 28,317 quality reads and 13,156 reads were classified by WIMP. The most abundant taxon was *H. sapiens* with 7,747 assigned reads, followed by *Candida* spp., accounting for 5,211 reads.

Sequencing of the 10^7 CFU spiked blood samples produced 17,238 quality reads, of which 15,988 reads were classified by WIMP. *H. sapiens* was the most abundant taxon with 15,934 reads and *Candida* spp. was the second most abundant taxon with 43 reads.

The 10^6 CFU spiked blood sample sequencing produced 501 reads that were analysed after passing the Albacore Q score filter and 339 reads were classified by WIMP. The most abundant taxon was *H. sapiens* with 339 assigned reads and *Candida* spp. was the third most abundant taxon with 3 reads.

Sequencing of the 10^5 CFU spiked blood samples generated 889,257 reads, of which 851,593 reads were successfully classified by WIMP. *H. sapiens* was the most abundant taxon with 851,512 reads, followed by *Candida* spp. which was assigned to 8 reads.

The 10^4 CFU spiked blood sample sequencing generated 10,423 reads that passed the Albacore Q score filter, of which 8,784 reads were classified by WIMP. The most abundant taxon was *H. sapiens* with 8,711 reads and *Candida* spp. Was the fourth most abundant taxon, accounting for 3 reads.

Sequencing runs in which *Candida* spp. Was the first or second most abundant taxon were 10^9 CFU, 10^8 CFU, 10^7 CFU and 10^5 CFU. As such, the limit of detection for saponin depleted samples was 10^5 CFU.

8.2.3.3 Percentage of *Candida* reads

The percentage of reads assigned to *Candida* spp. within reads that passed the Albacore Q score filter was calculated to compare the percentage in the sequencing of untreated samples and saponin depleted samples (Figure 1).

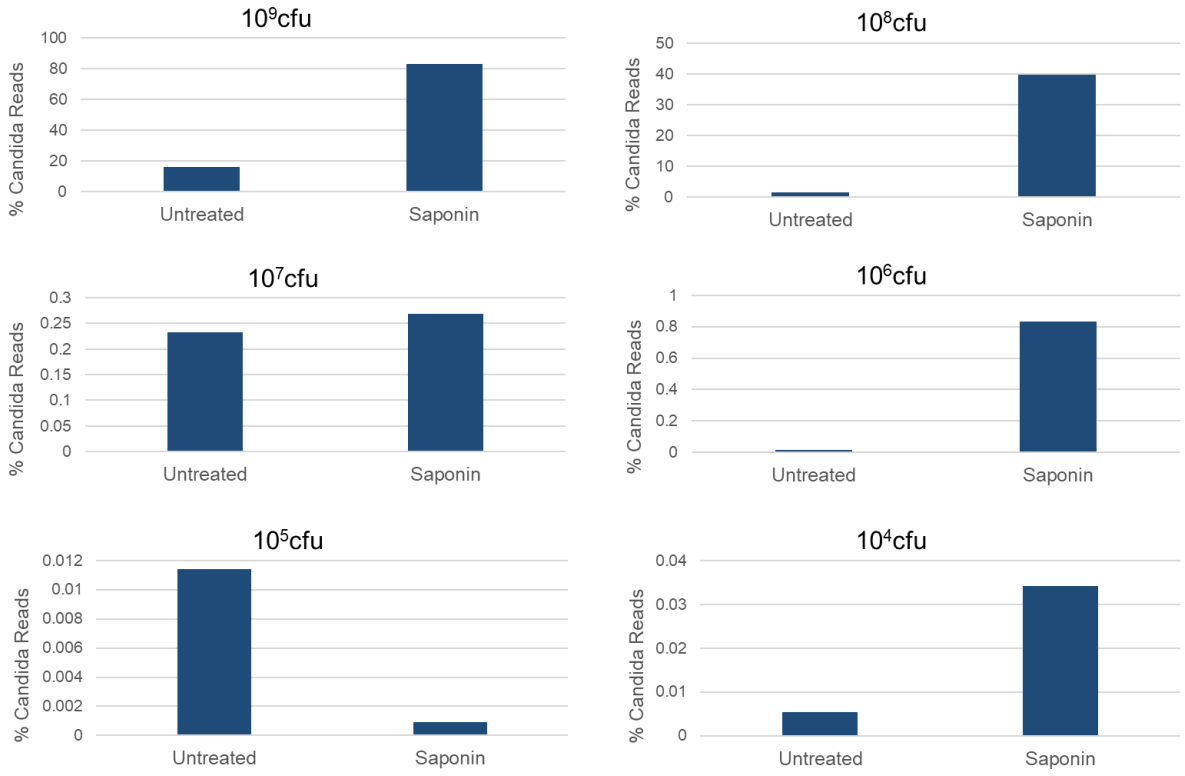


Figure 1: The percentage of reads assigned to *Candida* spp. after metagenomic nanopore sequencing of *C. albicans* spiked blood untreated and treated by human DNA saponin depletion.

For all concentrations except 10^5 CFU, the percentage of *Candida* spp. Reads was higher in samples that were treated with saponin depletion than those that were untreated. For 10^5 CFU

although the percentage of *Candida* spp. Reads was higher in the untreated condition, *Candida* spp. was the second most abundant taxon for the saponin depletion whereas it was the fifth most abundant taxon in the untreated condition.

8.3.3 Optimization of saponin human DNA depletion

The quantity of DNA before library preparation was measured by fluorescence with the DeNovix Broad Spectrum Kit and the quality was measured through A260/A230 and A260/A280 values using spectrophotometry. The yield of DNA after library sequencing that was loaded into the flow cell for sequencing was also measured by fluorescence. These values are recorded in Table 3.

Table 3: The quality and quantity of extracted DNA from *C. albicans* spiked blood from concentrations 10^7 CFU to 10^4 CFU when untreated and treated using saponin human depletion protocols.

		Before Library Preparation			
Host depletion method	***C. albicans* concentration (CFU)**	**DNA Yield (ng)**	**A260/A230**	**A260/A280**	**DNA loaded for sequencing (ng)**
Untreated	10^7	3542.4	2.17	1.98	910.0
	10^6	3970.0	2.20	1.99	1829.5
	10^5	4247.8	2.17	1.94	902.0
	10^4	5131.5	2.17	1.94	2506.0
Saponin Depletion Method 1	10^7	1272.5	0.55	1.65	121.0
	10^6	1632.5	0.61	1.64	173.0
	10^5	3025.0	1.04	1.82	794.0
	10^4	895.0	1.07	1.96	41.2
Saponin Depletion Method 2	10^7	4384.0	1.11	1.93	348.5
	10^6	4690.0	0.89	1.83	258.0
	10^5	3945.0	1.01	1.90	257.0

	10^4	X	X	X	X
Saponin Depletion Method 3	10^7	362.5	1.73	2.38	128.5
	10^6	344.0	1.40	1.97	97.5
	10^5	255.5	1.52	1.92	93.5
	10^4	201.0	1.54	1.79	90.5

The number of reads that successfully passed the Albacore Q score filter and were classified by WIMP for all sequencing runs are contained in Table 4. The number of reads classified as the taxa *H. sapiens* and *Candida* spp. are also recorded. For all runs *H. sapiens* was the most abundant taxa with the next most abundant taxa being *Candida* spp., aside from the saponin depletion method 1 10^5 CFU and untreated 10^4 CFU sequencing runs.

Table 4: The number of total classified reads, reads classified as *H. sapiens* and reads classified as *Candida* spp. of metagenomic sequencing of extracted DNA from C. albicans spiked blood from concentrations 10^7 CFU to 10^4 CFU when untreated and treated using saponin human depletion protocols.

***C. albicans* concentration (CFU)**	**Host depletion method**	**Total Classified Reads**	**Reads Classified *H. sapiens***	**Reads Classified *Candida* spp.**
10^7	Untreated	135,278	136,148	121
	Saponin Depletion Method 1	9,907	9,788	114
	Saponin Depletion Method 2	991,773	989,441	2,113
	Saponin Depletion Method 3	385,960	336,374	47,625
10^6	Untreated	144,551	144,543	4

	Saponin Depletion Method 1	34,896	34,863	27
	Saponin Depletion Method 2	1,209,671	1,209,148	344
	Saponin Depletion Method 3	194,518	191,813	2,277
10^5	Untreated	99,766	99,756	5
	Saponin Depletion Method 1	1,062,612	1,018,166	10
	Saponin Depletion Method 2	1,080,711	1,080,566	39
	Saponin Depletion Method 3	140,154	139,937	65
10^4	Untreated	980,419	980,354	3
	Saponin Depletion Method 1	154,949	149,842	1
	Saponin Depletion Method 3	96,191	96,058	3

To compare the ability of the human DNA depletion methods, the percentage of *Candida* spp. Assigned reads within the total classified reads was calculated (Figure 2). For all *C. albicans* spiked concentrations, aside from 10^5 CFU, the untreated condition produced the lowest percentage of *Candida* spp. assigned reads. For 10^5 CFU, saponin depletion methods 1 and 2 produced lower percentages of *Candida* spp. reads than the untreated condition. For all *C. albicans* spiked concentrations, saponin depletion method 3 produced the highest percentage of *Candida* spp. reads. For 10^7 CFU the *Candida* spp. Percentage of saponin depletion

method 3 (12.3%) was a 10.7-fold increase from the next highest *Candida* spp. Percentage from saponin depletion method 2 (1.2%). For 10^6 CFU, the *Candida* spp. Percentage of saponin depletion method 3 (1.17%) was a 15.1-fold increase from the next highest *Candida* spp. Percentage from saponin depletion 2 (0.08%). For 10^5 CFU, the *Candida* spp. Percentage of saponin depletion method 3 (0.046%) was a 9.3-fold increase from the next highest *Candida* spp. Percentage from the untreated condition (0.005%). For 10^4 CFU, the *Candida* spp. Percentage of saponin depletion method 3 (0.0031%) was a 4.8-fold increase from the next highest *Candida* spp. percentage from saponin depletion method 1 (0.0007%).

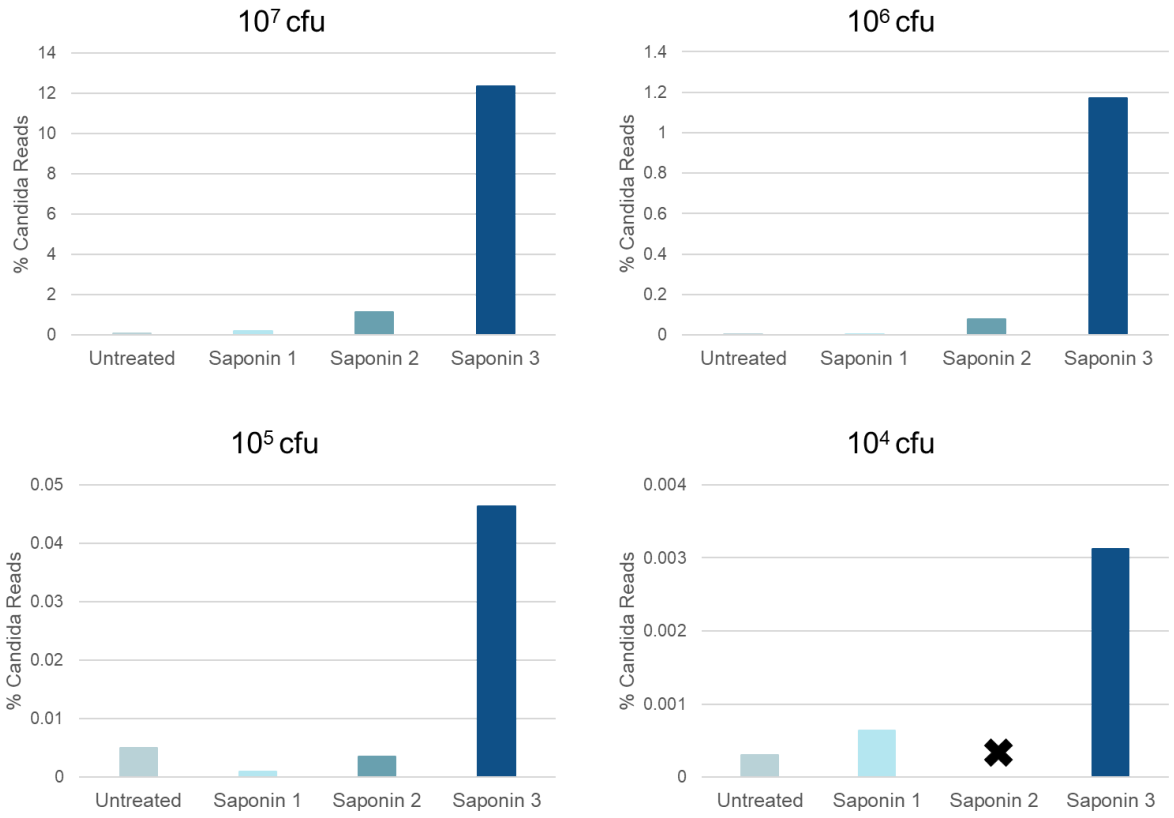


Figure 2: The percentage of *Candida* spp. Assigned reads after nanopore sequencing of untreated and saponin depletion treated *C. albicans* spiked blood at concentrations 10^7 CFU, 10^6 CFU, 10^5 CFU and 10^4 CFU. Saponin 1, saponin 2, and saponin 3 refer to three modifications of the saponin depletion protocol originally published by Charalampous *et al.* [21]. The sequencing run for the 10^4 CFU *C. albicans* spiked blood sample that underwent saponin depletion method 2 was unsuccessful and did not result in any analysed reads.

8.4 Discussion

Metagenomic sequencing with long-read sequencers is a promising approach to pathogen identification from clinical samples however many current limitations have been established. Metagenomic sequencing with long length sequencers is highly dependent upon the quantity and quality of DNA. There is no preliminary amplification to increase the amount of DNA being sequenced and as such, to produce accurate sequencing results, the DNA extraction methods must produce a high abundance of DNA. Nanopore sequencing requires pure DNA as sequencing pores can be blocked by contaminants, making the pore unable to sequence. The identification of pathogens from clinical samples is complicated by the high abundance of human DNA within samples and thus resulting in high amounts of human DNA sequenced and less pathogen DNA sequenced. This lowers the sensitivity of the test, and this is a prominent issue for metagenomic long-read sequencing as demonstrated previously in chapters 6 and 7.

The aim of this study was to select and optimize a human DNA depletion protocol to improve the identification of *C. albicans* in spiked blood samples using metagenomic sequencing with Flongle flow cells. Firstly, four depletion protocols were tested and compared to untreated DNA and a depletion protocol was selected based on the quantity and quality of DNA yield. The depletion protocol was then compared to untreated DNA extraction by identifying the limit of detection of *Candida* spp. Within spiked blood samples. The depletion protocol was then optimized to increase the percentage of *Candida* spp. reads sequenced in *C. albicans* spiked blood samples.

8.4.1 Comparison of human DNA depletion methods

Preliminary metagenomic sequencing studies have identified the need to reduce the proportion of host DNA within a sequencing sample. As such, protocols to improve the host

to pathogen ratio have been published alongside commercial kits aimed to produce the same result. These methods are aimed to deplete human DNA in a variety of clinical sample types such as saliva [20], blood [24], respiratory samples [21, 25, 26], tissue [22], and urine [27]. The methods also aim to deplete human DNA to enrich the presence of different microorganisms such as bacterial [21], fungal [28], viral [29], and parasitic [30] DNA.

For this study blood was selected as the clinical sample primarily due to its sterile nature and so when spiked and sequenced, only humans and the spiked species should be identified. Additionally, the ease of collection was considered in its selection.

Buffer RLT depletion method was previously used in the study by Rodriguez *et al.* to reduce lyse human cells from *Candida* spiked peripheral blood mononuclear cells [23]. This method relies upon Buffer RLT which is a reagent produced by QIAGEN to lyse human cells and tissue before RNA isolation [31, 32]. The additional inclusion of β-mercaptoethanol denatured RNAse within the sample. The final centrifugation and removal of supernatant would remove lysed human cells and contaminants while retaining fungal cells within the pellet. This method was targeted toward the preparation of samples to study the transcriptome and so was followed by RNA extraction in the initial study. However, this study aimed to extract DNA and so the resulting yield of DNA was too low to be measured and provided low A260/A230 values and inconsistent A260/A280 values, indicating high levels of contamination in the extracted DNA. As such, this method was not selected as the most effective human DNA depletion method.

The plasma depletion method was used by Gu *et al.* to deplete human DNA in clinical blood samples that were culture or PCR positive for pathogens [24]. The technique is reliant on the centrifugation of whole blood at a speed to separate human blood cells from microbial cells within plasma which are then removed. This technique resulted in DNA yield too low to be

detected and A260/A230 and A260/A280 values outside of the range that indicates high purity (Table 1). The centrifugation speed potentially caused the fungal cells to be in the lower layer and so were removed alongside the human cells, resulting in a low DNA yield. This method was not selected as the most effective depletion method due to the low DNA purity and yield.

The Molzym depletion method was the use of the commercial Molzym Ultra-Deep Microbiome Prep10 kit which provides human DNA depletion and extraction from many sample types such as blood and respiratory samples whilst retaining and enriching bacterial and fungal DNA [33]. Although the kit includes both depletion and extraction, in this study only the human DNA depletion portion of the protocol was applied to the spiked blood samples. The commercial kit has been used in other studies effectively to improve bacterial identification from clinical samples [34-36]. However, studies using commercial kits to enrich fungal DNA have not been published and its effectiveness is unverified. For its use in this study, the yield of DNA produced was too low to be measured and the resulting DNA had low purity and as such, this method was not selected as the most effective human DNA depletion method.

The saponin depletion method was a modified protocol from Charalampous *et al* which aimed to deplete human DNA within clinical respiratory samples to improve the identification of bacterial infections [21]. Saponin has been widely used for human cell lysis due to its efficient lysis through plasma membrane disruption and the minimal effects on bacteria and fungi [37, 38]. The osmotic shock from the addition of nuclease free water and salt (NaCl) is also seen in other saponin-based depletion methods to increase depletion [38, 39]. The method also had the inclusion of DNase to degrade and remove DNA which was not included in the other methods. The incubation at 37^0C was higher than incubation temperatures in the other methods and this is also reflected in other human DNA extraction

methods. Applying this method to blood presented challenges as the initial protocol included several centrifugation steps that removed reagents and contaminants however, when applied to blood, separation of layers or pelleting was not achieved. As such, the centrifuging steps were omitted which may have decreased the quality of DNA as contaminants were not removed.

This was determined to be the most successful human DNA depletion method as it produced the highest quantity and quality of DNA aside from the untreated condition (Table 1). Although the lower quantity of DNA may indicate more successful human DNA depletion, this yield is likely insufficient for library preparation and sequencing as there is no preamplification step and there is loss during library preparation. Further work can be done to improve the other human depletion methods such as optimizing the correct centrifuging speed in the plasma depletion method to successfully separate the fungal cells from the human cells. The combination of the human depletion and extraction steps for the Molzym Ultra-Deep Microbiome Prep10 kit may also be an alternative. Additionally, many other human depletion methods can be compared to find the optimal protocol but amongst those tested in this study, saponin was determined to be the most effective.

8.4.2 Limit of Detection

To determine whether human DNA depletion improved the host to pathogen ratio, the limit of detection of *C. albicans* spiked blood and untreated blood were investigated and compared. The limit of detection was determined to be the concentration in which reads assigned to *Candida* spp. Was the first or second most abundant taxon identified after sequencing. This is because blood should be sterile and free of other microbial reads aside from background reads, ideally resulting in only human and *Candida* spp. DNA being detected.

In the untreated condition, *Candida* spp. was the third most abundant taxa for the concentrations 10^9 CFU, 10^8 CFU, and 10^7 CFU. The inclusion of *E. coli* was excluded due to reasons stated earlier. At lower concentrations, *Candia* spp. Was fifth most abundant or less and so, the limit of detection for *Candida* spp. Spiked in untreated blood was 10^7 CFU.

The sequencing of saponin treated *C. albicans* spiked blood resulted in *Candida* spp. being the first or second most abundant taxa detected for 10^9 CFU, 10^8 CFU, and 10^7 CFU. For 10^6 CFU, *Candida* spp. Was the fifth most abundant taxa however, for 10^5 CFU *Candida* spp. Was the second most abundant taxa. The reason for this discrepancy is likely due to the low yield of the sequencing run, with only a total of 501 reads passing the quality control check. As the read count is so low it was unlikely to provide an accurate representation of sample DNA. With a more successful run, resulting in a higher number of successful reads, it is expected that *Candida* spp. Assigned reads would account for a larger proportion as reflected in the 10^5 CFU sequencing run. As such, the limit of detection of *C. albicans* for saponin treated spiked blood, was determined to be 10^5 CFU.

The limit of detection for *C. Albicans* in spiked blood was decreased from 10^7 CFU to 10^5 CFU when saponin depletion treatment was used before DNA extraction, thus improving the identification of *C. albicans* when using nanopore sequencing. This is also emphasised when comparing the percentage of *Candida* spp. reads in the total reads assigned (Figure 1). Saponin depleted samples outperformed the untreated samples in all concentrations except 10^5 CFU. This may be due to errors in the saponin depletion such as the removal of fungal cells during the initial centrifugation steps. At this concentration, there were 8 *Candida* spp. Assigned reads for both untreated and saponin treated, however, *Candida* spp. Was the second highest taxa in the saponin treated condition and fifth highest taxa in the untreated condition and so the depletion of human reads was inconclusive at this concentration.

The increase in the percentage of *Candida* spp. assigned reads for all other tested concentrations indicates a reduction in human DNA and alongside the improvement in the limit of detection, saponin treatment of clinical samples provides the potential to improve the identification of causative pathogens in sterile patient samples. This preliminary study also highlights the need for repeated testing to resolve the inconsistency seen within the 10^5 CFU concentration and optimisation of the saponin treatment protocol to further reduce the proportion of human reads to improve fungal identification.

8.4.3 Optimization of Saponin Depletion

Saponin has previously been determined to cause haemolysis of erythrocytes through cell membrane disruption [37, 40]. As such, it poses a potential for its use in human cell depletion, however, it has also been shown to have antifungal properties against the most common genera of fungal pathogens such as *Aspergillus*, *Candida*, *Cryptococcus*, and *Fusarium* [41]. The human DNA depletion method proposed by Charalampous *et al.* [21] implemented saponin to reduce the amount of human DNA in clinical samples for metagenomic long-read sequencing to identify bacterial samples. As such, the protocol required optimisation for effective application to fungal identification.

For all DNA extraction conditions, the amount of blood collected was increased from 200 µL, as in the limit of detection study, to 5 mL to increase the total number of reads and sensitivity of the protocol.

8.4.3.1 Saponin Depletion Method 1

The first saponin depletion method was similar to the initial saponin depletion method as described in section 8.2.1.2.2 however, an extended initial centrifugation step was included as the centrifugation speed was reduced due to limitations caused by the larger tubes

accommodating the larger sample volume. There was an improvement in the percentage of *Candida* spp. Classified reads for all concentrations except 10^5 CFU when compared to the untreated condition. Similarly, to the limit of detection study, this could be due to errors in saponin depletion as the yield of DNA and sequencing was high for both the untreated and saponin depletion method 1 sequencing runs. Although the *Candida* spp. Percentage was higher for the remaining concentrations, the difference was minimal.

8.4.3.2 Saponin Depletion Method 2

To account for the higher sample volume, the volumes for the nuclease free water and NaCl were increased. Also, the incubation times were increased to allow for more human DNA depletion. Additionally, a DNA purification step was included to improve the purity of sample DNA to improve sequencing with the nanopore which is highly reliant on pure sequencing material. The inclusion of this purification step improved the A260/A280 values.

The results of saponin depletion method 2 were similar to those in depletion method 1, as the *Candida* spp. Percentage was higher in all concentrations when compared to the untreated condition besides 10^5 CFU. DNA was extracted and purified for 10^4 CFU however, due to errors in loading the sequencing run was unsuccessful and did not produce any quality reads. Limitations in resources did not allow for this concentration to be repeated. Similarly to saponin depletion method 1, the improvement in *Candida* spp. percentage was minimal for the 10^7 CFU and 10^6 CFU. As such, further optimization was required.

8.4.3.3 Saponin Depletion Method 3

The changes made by the saponin depletion method were to increase all reagents, most notably the final saponin concentration was increased from 2.5% to 10% and the HL-SAN Buffer and HL-SAN DNase were doubled. Longer incubation times were also incorporated

into the protocol. The aim of these changes was for more haemolysis and the degradation of the resulting human DNA. To further improve the purity and sequencing of the samples, an additional purification step using AMPure beads was also added, similar to the steps included in library preparation. As such, the purity of the sample improved as demonstrated by the A260/A230 and A260/A280 values (Table 3).

As a result of these changes, there was at least a 15.1-fold to 4.8-fold increase in *Candida* spp. Percentage for all concentrations tested (Figure 2). Additionally, *Candida* spp. Was the first or second most abundant taxa for all concentrations and so the limit of detection has improved to 10^5 CFU as previously discussed in section 8.4.2. Although this is a large improvement, the percentage of *Candida* spp. assigned reads was still low, ranging from 12.4% - 0.003% and for higher confidence in results, the percentage would ideally be higher. Furthermore, the fungal load in blood stream infections is commonly below 10 CFU and it is estimated that 25% of clinical cases have a fungal load of 1 CFU or below [42, 43]. As such, *C. albicans* concentrations investigated in this study do not reflect the sensitivity required in clinical diagnosis, and further work is required to identify fungal pathogens at these concentrations.

The saponin depletion of human DNA should be further optimized to improve the proportion of pathogen DNA detected by long-read metagenomics. Higher concentrations of saponin should be tested to increase the lysis of human cells whilst retaining the fungal cells and avoiding its fungicidal properties. This may potentially be done through microscopy to observe the saponin concentration effect on human and fungal cells without requiring the expense of sequencing. Similarly, increasing the volume of HL-SAN Buffer and HL-SAN DNase should be investigated. This regent is highly valuable to the process due to its high concentration of DNase (25U/µl) whilst competitors are limited to 1-2U/µL. However, use in routine diagnostics may be challenging due to its high price point but bulk pricing is also

available and may bring costs down. Another avenue for optimization would be to test more DNA extraction methods such as commercial kits aimed at extraction from fungal cells. The current study focused on generating the highest yield of DNA after extraction and improving the ratio of the host to pathogen DNA through depletion methods in order to maximize sequencing yield. The implementation of a fungal focused extraction method may reduce the total DNA yield, but the resulting DNA may consist of a higher proportion of fungal DNA and so this should be further investigated.

8.4.4 Study Limitations

The current study presents preliminary research indicating the potential and basis for a workflow for the application of metagenomic long-read sequencing to the identification of fungal pathogens in sterile clinical samples. The major limitation of the study is the lack of repeats as each condition only resulted in one sequencing run however, there were no means for repetition due to a deficiency in resources. Further work will require repetition to resolve inconsistencies caused by human error or issues in sequencing which were highlighted throughout this study. The bioinformatics tools used within the study also limited the study as the WIMP bioinformatic tool characterised reads using fungal whole genomes which are unavailable for many fungal species and have minimal strain representation. For the identification of *C. albicans* in this study, WIMP was identified to the genus level using one *Candida glabrata* whole genome. This does not apply to a clinical setting as identification to the species level is essential to quickly inform on patient treatment due to inherent antifungal resistance that occurs in some fungal species. New fungal based bioinformatic tools would alleviate this in further work and these are discussed in chapters 5,6 and 9.

8.5 Conclusions

The inclusion of human DNA depletion techniques is essential to improve long-read metagenomic sequencing of clinical samples where pathogen load is low, such as in blood-borne fungal disease. Saponin has been demonstrated to be a promising reagent to deplete human DNA whilst retaining fungal DNA. A method to improve the limit of detection and host to pathogen ratio of *Candida* spp. in blood has been developed to be a basis for further work.

8.6 References

1. Magill, S.S., *et al.*, *Changes in prevalence of health care-associated infections in U.S. hospitals.* N. Engl. J. Med., 2018. **379**(18): p. 1732-1744.
2. Strollo, S., *et al.*, *Epidemiology of hospitalizations associated with invasive candidiasis, United States, 2002-2012.* Emerg. Infect. Dis., 2016. **23**(1): p. 7-13.
3. Pfaller, M.A., *et al.*, *Twenty years of the SENTRY Antifungal Surveillance Program: results for Candida species from 1997-2016.* Open Forum Infect. Dis., 2019. **6**(Suppl 1): p. S79-S94.
4. Colombo, A.L., J.N.A. Junior, and J. Guinea, *Emerging multidrug-resistant Candida species.* Curr. Opin. Infect. Dis., 2017. **30**(6): p. 528-538.
5. Toda, M., *et al.*, *Population-based active surveillance for culture-confirmed candidemia - Four sites, United States, 2012-2016.* MMWR Surveill. Summ., 2019. **68**(8): p. 1-15.
6. Pfaller, M., *et al.*, *Epidemiology and outcomes of candidemia in 3648 patients: data from the Prospective Antifungal Therapy (PATH Alliance(R)) registry, 2004-2008.* Diagn. Microbiol. Infect. Dis., 2012. **74**(4): p. 323-31.

7. Morrell, M., V.J. Fraser, and M.H. Kollef, *Delaying the empiric treatment of Candida bloodstream infection until positive blood culture results are obtained: a potential risk factor for hospital mortality.* Antimicrob. Agents Chemother., 2005. **49**(9): p. 3640-5.

8. Le Guern, R., *et al.*, *Comparison of time-to-positivity between two blood culture systems: a detailed analysis down to the genus-level.* Eur. J. Clin. Microbiol. Infect. Dis., 2021. **40**(7): p. 1399-1404.

9. Li, G., *et al.*, *Comparison of the performance of three blood culture systems in a Chinese tertiary-care hospital.* Front. Cell. Infect. Microbiol., 2019. **9**: p. 285.

10. Pfeiffer, C.D., *et al.*, *Quantitation of Candida CFU in initial positive blood cultures.* J. Clin. Microbiol., 2011. **49**(8): p. 2879-83.

11. Telenti, A., *et al.*, *Quantitative blood cultures in candidemia.* Mayo Clin Proc, 1991. **66**(11): p. 1120-3.

12. Clancy, C.J. and M.H. Nguyen, *Finding the "missing 50%" of invasive candidiasis: how nonculture diagnostics will improve understanding of disease spectrum and transform patient care.* Clin. Infect. Dis., 2013. **56**(9): p. 1284-92.

13. Schell, W.A., *et al.*, *Evaluation of a digital microfluidic real-time PCR platform to detect DNA of Candida albicans in blood.* Eur. J. Clin. Microbiol. Infect. Dis., 2012. **31**(9): p. 2237-45.

14. Ness, M.J., W.P. Vaughan, and G.L. Woods, *Candida antigen latex test for detection of invasive candidiasis in immunocompromised patients.* J. Infect. Dis., 1989. **159**(3): p. 495-502.

15. Nawrot, U., *et al.*, *Evaluation of blood culture media for the detection of fungi.* Eur. J. Clin. Microbiol. Infect. Dis., 2015. **34**(1): p. 161-167.

16. Tsang, C.C., *et al.*, *Rapid genomic diagnosis of fungal infections in the age of next-generation sequencing.* J. Fungi (Basel), 2021. **7**(8).

17. Tedersoo, L. and S. Anslan, *Towards PacBio-based pan-eukaryote metabarcoding using full-length ITS sequences.* Environ. Microbiol. Rep., 2019. **11**(5): p. 659-668.

18. Marquet, M., *et al.*, *Evaluation of microbiome enrichment and host DNA depletion in human vaginal samples using Oxford Nanopore's adaptive sequencing.* Sci. Rep., 2022. **12**(1): p. 4000.

19. Luth, T., *et al.*, *Improving analysis of the vaginal microbiota of women undergoing assisted reproduction using nanopore sequencing.* J. Assist. Reprod. Genet., 2022. **39**(11): p. 2659-2667.

20. Marotz, C., *et al.*, *Host DNA depletion in saliva samples for improved shotgun metagenomics.* Methods Mol. Biol., 2021. **2327**: p. 87-92.

21. Charalampous, T., *et al.*, *Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection.* Nat. Biotechnol., 2019. **37**(7): p. 783-792.

22. Heravi, F.S., *et al.*, *Host DNA depletion efficiency of microbiome DNA enrichment methods in infected tissue samples.* J. Microbiol. Methods., 2020. **170**: p. 105856.

23. Rodriguez, A., *et al.*, *Nucleic acids enrichment of fungal pathogens to study host-pathogen interactions.* Sci. Rep., 2019. **9**(1): p. 18037.

24. Gu, W., *et al.*, *Rapid pathogen detection by metagenomic next-generation sequencing of infected body fluids.* Nat. Med., 2021. **27**(1): p. 115-124.

25. Gan, M., *et al.*, *Combined nanopore adaptive sequencing and enzyme-based host depletion efficiently enriched microbial sequences and identified missing respiratory pathogens.* BMC Genomics, 2021. **22**(1): p. 732.

26. Shi, X., *et al.*, *Microfluidics-based enrichment and whole-genome amplification enable strain-level resolution for airway metagenomics.* mSystems, 2019. **4**(4).

27. Street, T.L., *et al.*, *Optimizing DNA extraction methods for nanopore sequencing of Neisseria gonorrhoeae directly from urine samples.* J. Clin. Microbiol., 2020. **58**(3).

28. Schwessinger, B. and J.P. Rathjen, *Extraction of high molecular weight DNA from fungal rust spores for long read sequencing.* Methods Mol. Biol., 2017. **1659**: p. 49-57.

29. Bal, A., *et al.*, *Quality control implementation for universal characterization of DNA and RNA viruses in clinical respiratory samples using single metagenomic next-generation sequencing workflow.* BMC Infect. Dis., 2018. **18**(1): p. 537.

30. Oyola, S.O., *et al.*, *Efficient depletion of host DNA contamination in malaria clinical sequencing.* J. Clin. Microbiol., 2013. **51**(3): p. 745-51.

31. Kasahara, T., *et al.*, *Evaluation of methods for duration of preservation of RNA quality in rat liver used for transcriptome analysis.* J. Toxicol. Sci., 2006. **31**(5): p. 509-519.

32. QIAGEN. *Buffer RLT*. 2022 [cited 2022; Available from: https://www.qiagen.com/us/products/discovery-and-translational-research/lab-essentials/buffers-reagents/buffer-rlt/.

33. Molzym. *Ultra-Deep microbiome prep*. 2022 [cited 2022; Available from: https://www.molzym.com/next-generation-sequencing/ultra-deep-microbiome-prep.

34. Chalupova, M., *et al.*, *Bacterial DNA detected on pathologically changed heart valves using 16S rRNA gene amplification.* Folia Microbiol. (Praha), 2018. **63**(6): p. 707-711.

35. Costantini, L., *et al.*, *Characterization of human breast tissue microbiota from core needle biopsies through the analysis of multi hypervariable 16S-rRNA gene regions.* Sci. Rep., 2018. **8**(1): p. 16893.

36. Leo, S., *et al.*, *Detection of bacterial pathogens from broncho-alveolar lavage by next-generation sequencing.* Int. J. Mol. Sci., 2017. **18**(9).

37. Bissinger, R., *et al.*, *Effect of saponin on erythrocytes.* Int. J. Hematol., 2014. **100**(1): p. 51-59.

38. Shi, Y., *et al.*, *Metagenomic sequencing for microbial DNA in human samples: emerging technological advances.* Int. J. Mol. Sci., 2022. **23**(4).

39. Zelenin, S., *et al.*, *Microfluidic-based bacteria isolation from whole blood for diagnostics of blood stream infection.* Methods Mol. Biol., 2017. **1547**: p. 175-186.

40. Baumann, E., *et al.*, *Hemolysis of human erythrocytes with saponin affects the membrane structure.* Acta. Histochem., 2000. **102**(1): p. 21-35.

41. Moghimipour, E. and S. Handali, *Saponin: Properties, methods of evaluation and applications.* Annu. Res. Rev. Biol., 2014. **5**(3): p. 207-220.

42. White, P.L., A. Shetty, and R.A. Barnes, *Detection of seven Candida species using the light-cycler system.* J. Med. Microbiol., 2003. **52**(Pt 3): p. 229-238.

43. Loeffler, J., *et al.*, *Quantification of fungal DNA by using fluorescence resonance energy transfer and the light cycler system.* J. Clin. Microbiol., 2000. **38**(2): p. 586-90.

Chapter 9
Bioinformatic Improvements in Fungal Metagenomics - Inferring Species Compositions of Complex Fungal Communities from Long- and Short-Read Sequence Data

9.1 Inferring Species Compositions of Complex Fungal Communities from Long- and Short-Read Sequence Data

This manuscript was published in mBio on the 11th April 2022 (volume 12, issue 2, pages 1-16). Hoang M.T.V. was involved in sample preparation, generation of data and manuscript preparation. Hu Y. is the primary author and was responsible for sample preparation, data generation, bioinformatic analysis and manuscript preparation. Irinyi L. was involved in sample preparation, data generation and manuscript preparation. Eenjes T. and Graetz A. performed bioinformatic analysis and Stone E. provided feedback on statistical analysis. Meyer W., Stone E., Schwessinger B., and Rathjen J.P. conceived the study and designed experiments. All authors contributed to the data analysis and preparation of the manuscript. All authors were involved in the revision of the manuscript and support its inclusion in this thesis.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Minh Thuy Vi Hoang
30th June 2023

As a supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Professor Wieland Meyer
30th June 2023

AMERICAN SOCIETY FOR MICROBIOLOGY | mBio®

RESEARCH ARTICLE

Inferring Species Compositions of Complex Fungal Communities from Long- and Short-Read Sequence Data

Yiheng Hu,[a*] Laszlo Irinyi,[b,c,d] Minh Thuy Vi Hoang,[b,c,d] Tavish Eenjes,[a] Abigail Graetz,[a] Eric A. Stone,[a,e] Wieland Meyer,[b,c,d,f,g] Benjamin Schwessinger,[a] John P. Rathjen[a]

[a]Research School of Biology, Australian National University, Canberra, ACT, Australia
[b]Molecular Mycology Research Laboratory, Centre for Infectious Diseases and Microbiology, Faculty of Medicine and Health, Sydney Medical School, Westmead Clinical School, The University of Sydney, Sydney, NSW, Australia
[c]Sydney Institute for Infectious Diseases, The University of Sydney, Sydney, NSW, Australia
[d]Westmead Institute for Medical Research, Westmead, NSW Australia
[e]ANU-CSIRO Centre for Genomics, Metabolomics and Bioinformatics, Canberra, ACT, Australia
[f]Westmead Hospital (Research and Education Network), Westmead, NSW, Australia
[g]Curtin Medical School, Curtin University, Perth, Bentley, WA, Australia

ABSTRACT The kingdom Fungi is highly diverse in morphology and ecosystem function. Yet fungi are challenging to characterize as they can be difficult to culture and morphologically indistinct. Overall, their description and analysis lag far behind other microbes such as bacteria. Classification of species via high-throughput sequencing is increasingly becoming the norm for pathogen detection, microbiome studies, and environmental monitoring. With the rapid development of sequencing technologies, however, standardized procedures for taxonomic assignment of long sequence reads have not yet been well established. Focusing on nanopore sequencing technology, we compared classification and community composition analysis pipelines using shotgun and amplicon sequencing data generated from mock communities comprising 43 fungal species. We show that regardless of the sequencing methodology used, the highest accuracy of species identification was achieved by sequence alignment against a fungal-specific database. During the assessment of classification algorithms, we found that applying cutoffs to the query coverage of each read or contig significantly improved the classification accuracy and community composition analysis without major data loss. We also generated draft genome assemblies for three fungal species from nanopore data which were absent from genome databases. Our study improves sequence-based classification and estimation of relative sequence abundance using real fungal community data and provides a practical guide for the design of metagenomics analyses focusing on fungi.

IMPORTANCE Our study is unique in that it provides an in-depth comparative study of a real-life complex fungal community analyzed with multiple long- and short-read sequencing approaches. These technologies and their application are currently of great interest to diverse biologists as they seek to characterize the community compositions of microbiomes. Although great progress has been made on bacterial community compositions, microbial eukaryotes such as fungi clearly lag behind. Our study provides a detailed breakdown of strategies to improve species identification with immediate relevance to real-world studies. We find that real-life data sets do not always behave as expected, distinct from reports based on simulated data sets.

KEYWORDS bioinformatics, fungi, metagenomics, pathogens

Fungi are ubiquitous yet their presence and impact are often overlooked. It has been estimated that 2.2–3.8 million fungal species inhabit planet earth (1) but only about 4% of these are catalogued (2). Fungi play diverse and fundamental roles throughout

Editor B. Gillian Turgeon, Cornell University

Copyright © 2022 Hu et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Address correspondence to Yiheng Hu, yiheng.hu@anu.edu.au, Benjamin Schwessinger, benjamin.schwessinger@anu.edu.au, or John P. Rathjen, john.rathjen@anu.edu.au.

*Present address: Yiheng Hu, Department of Microbial Interactions, IMIT/ZMBP, University of Tübingen, Tübingen, Germany.

The authors declare no conflict of interest.

Received 15 August 2021
Accepted 28 February 2022
Published 11 April 2022

Downloaded from https://journals.asm.org/journal/mbio on 08 June 2022 by 54.66.212.33.

evolution. For example, saprotrophs break down dead organic matter to release nutrients, whereas mycorrhizae facilitate nutrient uptake by plants in soil. Fungi constitute a major disease load to humans, causing millions of deaths per year, and wreak devastating crop losses via a constant toll of disease and are an existential threat to many plant species (3, 4). On the other hand, fungi are or are used to manufacture delicious foods and beverages and have saved countless lives via antibiotic production (5, 6). Therefore, a call was made recently to expand fungal research and improve our awareness of this special kingdom (7).

To progress our understanding of fungal biology we need to improve our ability to correctly identify species from complex mycological samples. Historically, taxonomic assignment was based on morphological and reproductive traits, but this has been largely surpassed by DNA-based classification which revolutionized mycology, not only refining the conventional taxonomic tree (8, 9) but also standardizing the identification of species. In the absence of whole-genome data, DNA-based classification primarily exploits the internal transcribed spacer (ITS) within the rRNA genes as a highly polymorphic marker to distinguish species, being designated the primary fungal barcode (10). It is easily amplified and sequenced due to highly conserved flanking sequences and contains a high degree of variation between even closely related species. Although a mature pipeline comprising ITS amplification, Illumina sequencing and data analysis has been established (11), several studies reported biases from the sequencing technology used and from unevenly amplified fungal marker regions (12–14). Recently, novel strategies exploiting long-range amplification and long-read sequencing have been developed to improve these classifications (15, 16). In addition, whole-genome shotgun sequencing and rapidly expanding genome databases allow mapping of newly generated DNA sequences directly to the database. This so-called shotgun metagenomics strategy has now been applied widely to dissect the structure and composition of complex microbial communities. Shotgun metagenomics also allows exploitation of genetic variation throughout the genome and abandonment of the marker gene amplification step, which increases classification accuracy and reduces the biases in estimation of relative abundance (17).

Although advanced sequencing methods allow novel strategies for fungal identification particularly from mixed samples, new demands are placed on data analysis pipelines to improve the accuracy of sequence classification. Various algorithms have been developed to classify DNA sequences at distinct taxonomic ranks based on sequence databases containing taxonomic information (18–22). For example, alignment algorithms such as Basic Local Alignment Search Tool (BLAST) (19) detect matches of each sequence to subjects of the target database along with the taxonomic information assigned to each entry. Alternatively, sequence features represented by short unique subsequences named k-mers can be derived from sequence data and mapped to databases to identify taxa with the highest number of cross-mapping k-mers (20). Several studies have critically assessed algorithms for species classification on simulated data sets or bacterial community data sets (23–25), but benchmarking of long read sequencing strategies for taxonomic classification of fungal communities using real data are extremely rare. As such, the bioinformatic tools to process fungal sequencing data, particularly long reads, are not as well developed as those for bacteria. In addition to bioinformatic tools, the choice of database also influences classification dramatically, but only a few studies have researched their impact (26–28). Compared to bacterial genomes, fungi contain complex genetic features such as multiple chromosomes, expanded repeat regions and larger genome sizes, all of which introduce inaccuracies during sequence classification. Therefore, comprehensive benchmarking of both classification algorithms and databases is needed to optimize identification pipelines for the kingdom fungi.

Here, we assessed different combinations of algorithms and databases during processing of both short- and long-read sequencing data for the identification of taxa from mock fungal communities. We constructed a low diversity community consisting of ascomycetous and basidiomycetous fungal species, which were designed to mimic fungal identification scenarios from clinical or other host-associated environmental

Downloaded from https://journals.asm.org/journal/mbio on 08 June 2022 by 54.66.212.33.

samples. We identified key factors that influence the accuracy of classifications, both for mock community data sets and public data sets. Optimization of these methods lead to more accurate community compositions. Our results provide guidelines for the design of sequence-based community analysis for fungal species.

RESULTS

Construction of mock fungal community data sets. We constructed two mock communities from the same set of 43 fungal species (Table 1), consisting of yeasts and filamentous human-associated pathogens. We aimed to mimic a real-world situation with a reasonable level of species diversity in order to adjust different technical search parameters while aiming to recover all species. One community comprised pooled DNA individually extracted from each species and the second was composed of DNA extracted from equal quantities of fungal biomass (pooled biomass) of each species mixed together prior to the extraction. We generated four sequence data sets for each community using Illumina and nanopore technologies, sequencing both shotgun metagenomes and targeted amplicons, respectively. The data derived from each strategy are summarized in Table 2.

Alignment algorithm against a fungal-specific database resulted in the most accurate fungal classification. We compared different analysis strategies for each metagenomics shotgun data set. For nanopore data sets, we directly used the quality-controlled reads for classification. For Illumina data, we quality filtered all reads and assembled them into contigs before classification to maximize the classification accuracy. We performed both alignment and k-mer based classifications on these data using BLAST and Kraken2 (19, 21) using a "winner takes all" strategy in which the top hit was taken to assign the taxonomic classification of the query sequence. For each algorithm, we compared the use of two reference databases: the nonredundant NCBI nucleotide database (29) and the RefSeq fungi database (30) which only contains curated fungal genomes. We first assessed the performance of each alignment tool on both databases for each data input. We compared the concordance in the results of each pipeline at the genus level. We define concordance as the percentage of fungal genera identified by both analyses in a pairwise comparison (Fig. 1A). The concordance between analyses on each data set varied between 69% and 86% and generally, Illumina data resulted in a higher concordance than the nanopore data.

We then aimed to identify the combination of algorithm and database that yielded the most accurate species identification. We used classification proportion and precision to evaluate each classification, where Classification Proportion $= \frac{\#\ total\ basepairs\ classified.}{\#\ total\ basepairs\ of\ input\ reads}$, and Precision $= \frac{\#\ total\ basepairs\ classified\ correctly}{\#\ total\ basepairs\ classified}$.

The number of total base pairs is calculated as total read length for nanopore reads and total coverage of Illumina reads to each contig (24, 25). We plotted the precision and classification proportion for each pipeline and found three regular patterns (Fig. 1B and C): First, for each data set, BLAST resulted in higher precision but lower classification proportion by comparison to Kraken2. Second, Illumina contigs returned a higher classification proportion and precision than nanopore reads. Third, classification against the Refseq fungal database yielded higher precision than those against the NCBI nucleotide database. In summary, BLAST alignments against the Refseq fungal database yielded the best classification strategy for shotgun metagenomics data sets. Interestingly, this result contrasts to the method of choice for amplicon data sets, in which the use of a restricted database generates higher false positives (27, 31).

Applying cutoffs to query coverage improves classification accuracy on shotgun metagenomics data sets. We next aimed to improve our classification scheme by filtering the BLAST search output. We reasoned that restricting alignment metrics would reduce the number of false classifications. To investigate changes in classification accuracy after restricting BLAST output parameters, we first BLASTed shotgun metagenomics reads against the RefSeq fungi database without applying any filter, then applied progressive cutoffs on different parameters of the BLAST results. We evaluated changes in

Downloaded from https://journals.asm.org/journal/mbio on 08 June 2022 by 54.66.212.33.

TABLE 1 Metadata of the mock fungal community

Species name	Strain used in the mock community	Corresponding reference genome strain	Assembly level	NCBI accession	File name
Aspergillus flavus	WM 03.225	NRRL3357	Scaffold	GCF_000006275.2	GCF_000006275.2_JCVI-afl1-v2.0_genomic.fna.gz
Aspergillus fumigatus	WM 06.98	Af293	Chromosome	GCF_000002655.1	GCF_000002655.1_ASM265v1_genomic.fna.gz
Blastobotrys proliferans	WM 07.12 (CBS 522.75 Type strain)	NRRL Y-17577	Scaffold	GCA_003707485.2	GCA_003707485.2_ASM3707485v2_genomic.fna.gz
Candida albicans	WM 229 (CBS 562 Type strain)	SC5314	Chromosome	GCA_000182965.3	GCA_000182965.3_ASM18296v3_genomic.fna.gz
Candida dubliniensis	WM 02.73	CD36	Chromosome	GCA_000026945.1	GCA_000026945.1_ASM2694v1_genomic.fna.gz
Candida glabrata	WM 03.500	CBS 138	Chromosome	GCF_000002545.3	GCF_000002545.3_ASM254v2_genomic.fna.gz
Candida haemulonis	WM 890	B11899	Contig	GCF_002926055.2	GCF_002926055.2_CanHae_1.0_genomic.fna.gz
Candida parapsilosis	WM 02.200	CDC317	Scaffold	GCA_000182765.2	GCA_000182765.2_ASM18276v2_genomic.fna.gz
Candida tropicalis	WM 01.203	MYA-3404	Scaffold	GCF_000006335.3	GCF_000006335.3_ASM633v3_genomic.fna.gz
Clavispora lusitaniae	WM 18 (CBS 4413 Type strain)	ATCC 42720	Scaffold	GCF_000003835.1	GCF_000003835.1_ASM383v1_genomic.fna.gz
Naganishia albida (former Cryptococcus albidus)	WM 773 (CBS 142 Type strain)	JCM 2334	Scaffold	GCA_001599735.1	GCA_001599735.1_JCM_2334_assembly_v001_genomic.fna.gz
Cryptococcus gattii VGI	WM 179 (VGI Standard strain)	WM 276	Chromosome	GCF_000185945.1	GCA_000185945.1_ASM18594v1_genomic.fna.gz
Cryptococcus gattii VGII	WM 178 (VGII Standard strain)	R265	Scaffold	GCA_000786445.1	GCA_000786445.1_R265.1_genomic.fna.gz
Cryptococcus gattii VGIII[a]	WM 175 (VGIII Standard strain)				
Cryptococcus gattii VGIV[a]	WM 779 (VGIV Standard strain)				
Filobasidium magnum (former Cryptococcus magnus)[b]	WM 13.104				
Cryptococcus neoformans VNI	WM 148 (VNI Standard strain)	JEC21	Chromosome	GCA_000091045.1	GCA_000091045.1_ASM9104v1_genomic.fna.gz
Cryptococcus neoformans VNII	WM 626 (VNII Standard strain)	B-3501A	Chromosome	GCF_000149385.1	GCF_000149385.1_ASM14938v1_genomic.fna.gz
Cryptococcus neoformans VNIV	WM 629 (VNIV Standard strain)	H99	Chromosome	GCF_000149245.1	GCF_000149245.1_CNA3_genomic.fna.gz
Cyberlindnera jadinii	WM 45 (CBS 1600 Type strain)	NBRC0988	Chromosome	GCA_000328385.1	GCA_000328385.1_Cutl_1.0_genomic.fna.gz
Debaryomyces hansenii	WM 309 (CBS 767 Type strain)	CBS767	Chromosome	GCF_000006445.2	GCF_000006445.2_ASM644v2_genomic.fna.gz
Diutina catenulata (former Candida catenulata)[b]	WM 03.477	WY3-10-4	Contig	GCA_003285555.1	GCA_003285555.1_ASM328555v1_genomic.fna.gz
Diutina mesorugosa (former Candida mesorugosa)[b]	WM 03.468				
Diutina rugosa (former Candida rugosa)[b]	WM 03.463				
Geotrichum candidum	WM 05.217	CLIB 918	Scaffold	GCA_001402995.1	GCA_001402995.1_New2.3_08062011_genomic.fna.gz
Geotrichum fermentans	WM 17.18	CICC 1368	Scaffold	GCA_002233575.1	GCA_002233575.1_ASM223357v1_genomic.fna.gz
Kluyveromyces marxianus	WM 13 (CBS 834)	DMKU3-1042	Complete Genome	GCF_001417885.1	GCF_001417885.1_Kmar_1.0_genomic.fna.gz
Kodamaea ohmeri	WM 10.200	148	Contig	GCA_004919595.1	GCA_004919595.1_Kodohm_148_genomic.fna.gz
Lomentospora prolificans	WM 13.369	JHH-5317	Contig	GCA_002276285.1	GCA_002276285.1_Lprolificans_pilon_genomic.fna.gz
Meyerozyma caribbica	WM 03.389	MG20W	Contig	GCA_000755205.1	GCA_000755205.1_ASM75520v1_genomic.fna.gz
Meyerozyma guilliermondii	WM 02.131	ATCC 6260	Scaffold	GCF_000149425.1	GCF_000149425.1_ASM14942v1_genomic.fna.gz
Pichia kudriavzevii	WM 02.78	CBS573	Complete Genome	GCA_003054445.1	GCA_003054445.1_ASM305444v1_genomic.fna.gz
Pichia membranifaciens	WM 32 (CBS 638 Type strain)	NRRL Y-2026	Scaffold	GCF_001661235.1	GCF_001661235.1_Picme2_genomic.fna.gz
Pichia norvegensis	WM 885	NRRL Y-7687	Scaffold	GCA_003705465.1	GCA_003705465.1_ASM370546v1_genomic.fna.gz
Purpureocillium lilacinum[c]		PLFJ-1		GCA_004026455.1	GCA_004026455.1_ASM402645v1_genomic.fna
Rhodotorula mucilaginosa	WM 09.204	RIT389	Contig	GCA_002250355.1	GCA_002250355.1_RIT389_v1_genomic.fna.gz
Scedosporium aurantiacum	WM 06.385	WM 09.24	Scaffold	GCA_000812075.1	GCA_000812075.1_ASM81207v1_genomic.fna.gz
Scedosporium boydii (former Pseudallescheria boydii)	WM 09.122 (CBS 101.22 Type strain)	IHEM 23826	Contig	GCA_002221725.1	GCA_002221725.1_ScBoyd1.0_genomic.fna.gz
Trichophyton rubrum	WM 04.474	CBS 118892	Scaffold	GCF_000151425.1	GCF_000151425.1_ASM15142v1_genomic.fna.gz
Trichosporon asahii	WM 03.423	CBS 2479	Scaffold	GCF_000293215.1	GCF_000293215.1_Trichosporon_asahii_1_genomic.fna.gz
Cutaneotrichosporon dermatis (former Trichosporon dermatis)	WM 601	JCM 11170	Scaffold	GCA_003116895.1	GCA_003116895.1_JCM_11170_assembly_v001_genomic.fna.gz
Wickerhamomyces anomalus	WM 03.507	NRRL Y-366-8	Scaffold	GCF_001661255.1	GCF_001661255.1_Wican1_genomic.fna.gz
Yarrowia lipolytica	WM 17 (CBS 6124 Type strain)	CLIB122	Chromosome	GCF_000002525.2	GCF_000002525.2_ASM252v1_genomic.fna.gz
Zygoascus hellenicus[d]	WM 02.460 (CBS 5839 Type strain)	Y-7136	Scaffold		Zyghe1_2_AssemblyScaffolds_Repeatmasked.fasta

[a]No publicly available reference genome, but high quality reference genomes of relative strains.
[b]Locally sequenced and assembled.
[c]Potential contamination species.
[d]Downloaded from JGI, no NCBI accession.

Downloaded from https://journals.asm.org/journal/mbio on 08 June 2022 by 54.66.212.33.

TABLE 2 Characteristics of each sequence data set

Sample	Sequencing tech	Sequencing strategy	Number base pairs	Number reads	Number assembled contigs	Number mapped base pairs (Gb)
Pooled DNA	Illumina	Shotgun	3.91 Gb	14,525,058	338,823	3.69
		Amplicon	66.9/95.8/106.4 Mb[a]	39,374/9,614/10,236[b]	NA	N/A
	Nanopore	Shotgun	1.96 Gb	1,273,484	NA	N/A
		Amplicon	71.5/72.5/86.5 Mb	26,212/ 26,680/ 31,826[b]	NA	N/A
Pooled biomass	Illumina	Shotgun	3.67 Gb	13,623,120	345,009	3.44
		Amplicon	55.7/38.1/71.9 Mb[a]	23,613/13,828/27,093[b]	NA	N/A
	Nanopore	Shotgun	3.78 Gb	1,043,343	NA	N/A
		Amplicon	54.5/49.4/42.0 Mb	20,163/ 18,273/ 15,502[b]	NA	N/A

[a]The total number of basepairs of each technical replicate was calculated before import into QIIME2 pipeline.
[b]Number of nanopore reads or paired-end Illumina reads for technical replicate 1/replicate 2/replicate 3 after quality control.

the results based on the three metrics: precision, remaining rate, and completeness. Precision is described above and estimates the accuracy of the classification; remaining rate captures the percentage of the input data remaining after the application of each cutoff; and completeness is the number of taxa captured relative to the total number of taxa within the mock community. We initially applied cutoffs on query length; E value - the number of expected hits of similar quality that could be found by chance alone; and pident——the percentage of identical matches within the region of alignment between query and subject. As shown in Fig. 2A, applying progressive cutoffs to query length did not improve the precision, while both completeness and remaining rate diminished dramatically from very small cutoff values. Cutoffs applied to alignment E values removed <20% of the BLAST results, whereas precision showed minor improvement, especially on nanopore data sets (Fig. 2B). For Illumina data, applying cutoffs to the E value increased the precision by around 2% but at the cost of diminished completeness. E value cutoffs performed better on nanopore data sets, improving precision by 3% (pooled DNA) or 4% (pooled biomass) with non-identification of only a single genus from the mock community, at 10^{-250} or almost 10^{-400}, respectively. Progressive cutoffs on pident yielded the best results of all three filters. For Illumina data, precision was improved by up to 8% for data from pooled biomass sample, and completeness remained at 100% in almost all cases (Fig. 2C). For nanopore data sets, pident cutoffs improved the precision by up to ~3% before sharp decreases with a concurrent filtering of ~60% BLAST result as shown by the remaining rate. We think that these sharp decreases of precision are correlated with the nanopore error rate of 10% at the time. Given the characteristically high error rate of nanopore reads, we also applied cutoffs on quality scores to these data. Cutoffs applied to Phred scores did not alter the precision, while a significant proportion of the data set was lost through filtering (Fig. S1). Overall, our results suggest that applying E value and pident filters to BLAST results perform well

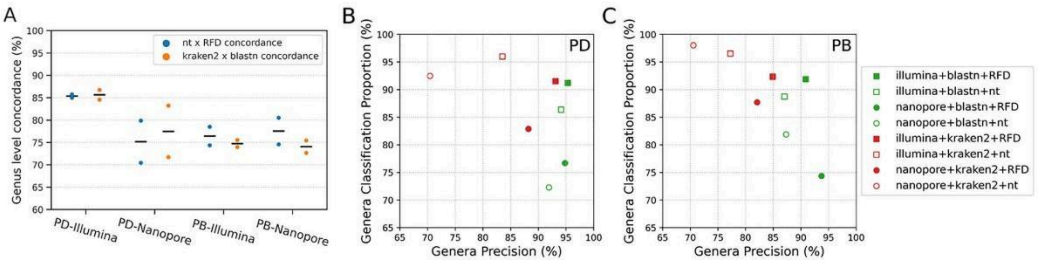


FIG 1 Analysis of shotgun metagenomics data. (A) Swarmplot showing the concordance in genus identification after varying either the alignment algorithm or querying different databases on different data inputs. nt = NCBI nucleotide database (29); RFD = RefSeq Fungi database (30); data inputs are indicated below the line (PD = pooled DNA; PB = pooled biomass); (B) Identification of fungal genera from PD samples. The classification proportion and precision were derived from different combinations of search algorithms and databases as indicated (box); (C) Identification of fungal genera from pooled biomass samples. The classification proportion and precision were derived from the different combinations of search algorithms and databases as indicated.

Downloaded from https://journals.asm.org/journal/mbio on 08 June 2022 by 54.66.212.33.

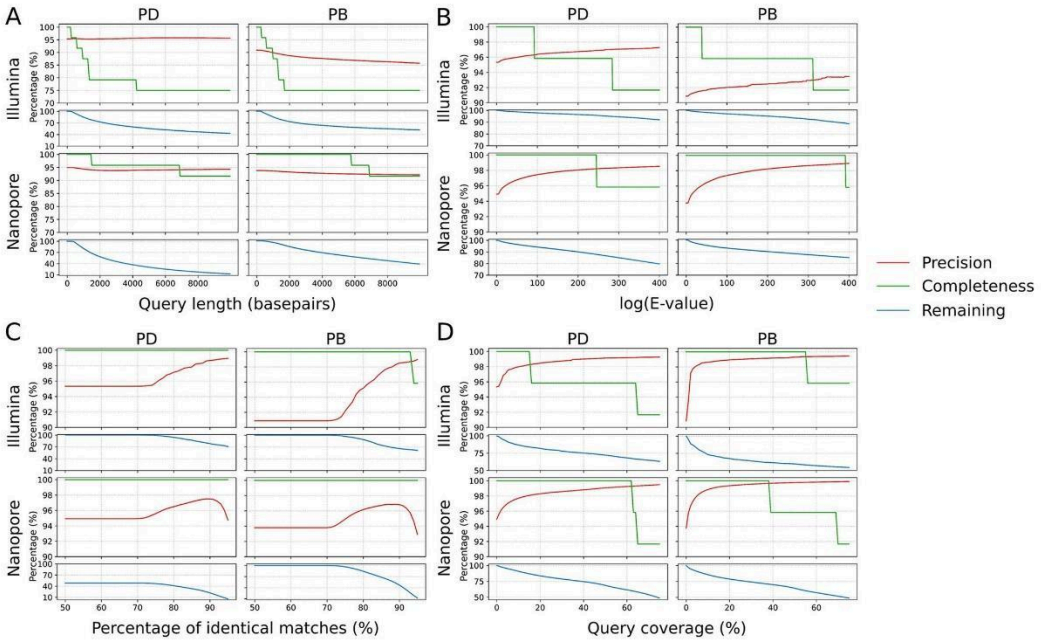


FIG 2 Dynamics in precision, completeness, and remaining rate after applying progressive cutoffs on BLAST alignment metrics. PD = pooled DNA; PB = pooled biomass. (A) Cutoffs applied to query length. (B) cutoffs applied to alignment E values. (C) cutoffs applied to the percentage of identical matches. (D) cutoffs applied to query coverage.

on either Illumina or nanopore data but not both, and that cutoffs based on query length or quality scores did not affect the precision significantly.

Given the results above, we investigated how the alignment parameters were calculated and explored other variables to improve the classifications. The BLAST E value is calculated as $E = mn2^{-S}$ in which S is the bits score derived from the number of gaps and mismatches in the alignment, and m and n are the query length and database total length, respectively (32). Therefore, the E value is influenced exponentially by the alignment quality. We next investigated query coverage, a metric based on how much of the query sequence aligned to the subject. We calculated the query coverage as the number of identical matches divided by the read or contig length and applied progressive cutoffs on this parameter for each data set/algorithm analysis. As shown in Fig. 2D, applying cutoffs on query coverage improved the precision of all four analyses significantly, and did not cause losses of completeness at smaller cutoff values. For example, at a 10% cutoff on query coverage, the precision of all four analyses was 98–99% while the completeness remained at 100% and the remaining rate dropped by 10–25%. This result not only supported our hypothesis that the total length of the alignment matters as much as the alignment quality, but also suggested a novel approach to improve the accuracy of fungal classification.

Improving taxa identification from published metagenomics data sets using query coverage as a filtering parameter. After improving classifications by applying cutoffs to the query coverage on the mock community data sets, we extended this strategy to try to improve the classification of published shotgun metagenomics data sets. We reanalyzed 10 nanopore and six Illumina shotgun metagenomics data sets (33–36). These included host-associated fungal samples (nanopore) and host-depleted microbiome data (Illumina). For unified comparison, we used the BLAST algorithm for all samples, and the choice of database was primarily based on the specific aim of each study. For example, for all Illumina data sets, we downloaded the quality-controlled

Downloaded from https://journals.asm.org/journal/mbio on 08 June 2022 by 54.66.212.33.

TABLE 3 Assignment of published sequence data to genera after application of cutoffs to query coverage

Sample ID	Sample description	Sequencing tech	Cutoffs on query coverage (%)	Filtered results (%)	Percentage of confirmed genera before applying cutoffs (%)	Percentage of confirmed genera after applying cutoffs (%)
a1	Human sputum samples[35]	Nanopore	59	20.2	85.9	86.5
a2			53.2	20.1	97.9	98.5
a3			54	20.5	96.5	97.4
a4			45.5	20.1	16.2	19.8
a5			58.5	20	71.1	66.9
a6			50.4	20.1	93.6	94.7
b1	Field infected wheat samples[34]		5	20	60.4	75.1
b2			0.77	19.9	34.8	43
b3			12	19.7	67	82
b4			0.61	20	5.8	6.2
c1	Swine gut microbiome samples[38]	Illumina	2.4	20.1	32	35.4
c2			3.3	20.2	34.2	36.6
c3			2.6	20.2	35.2	38.3
d1	Mouse gut microbiome samples[39]		3.4	19.8	29.1	24.3
d2			14	20.1	63.7	69.4
d3			4.5	20.2	38.6	42.3

sequences and reanalyzed them using the assembly and BLAST pipeline described above against the NCBI nucleotide database since they investigated the total microbiome diversity of all species. For the nanopore human data sets (35), we used the BLAST results taken directly from the original articles for analysis focusing on the NCBI nucleotide database. For the infected wheat data sets (34), we downloaded the sequences and reanalyzed them against the RefSeq fungal database as the aim of this study is to identify a limited number of fungal pathogen.

Since the environmental data sets contain unknown species, we followed the concept of precision to evaluate the classification of each sample. We calculated the percentage of the data set that was classified into taxa known to be included in the sample. For example, in reanalyzing human clinical samples (35), we included the pathogen (*Pneumocystis jirovecii*) and the human host (*Homo sapiens*) as the known taxa and calculated the total proportion of query sequences classified to these taxa before and after applying cutoffs on query coverage. Table 3 shows the improvement in taxonomic classification from the published data sets after applying query coverage cutoffs. We initially applied a 20% cut-off on the query coverage for all analyses, but the data loss in most cases was too high. Therefore, we applied query cutoffs that filtered around 20% of the BLAST results (20% drop of remaining rate) based on our analysis of the mock fungal community data sets (Fig. 2D). The percentage of confirmed genera increased for nearly all data sets after applying query coverage cutoffs (Table 3). For the Illumina microbiome data sets, we first assessed the change of proportions in fungal taxa after applying cutoffs on query coverages using the species lists identified by Donovan et al. (37) as confirmed taxa. We observed only marginal increase in percentages for the confirmed fungal communities, due to their low total proportions in the original samples. We then calculated the improvement in precision for the bacterial communities. The Illumina data sets were generated from swine and mouse gut microbiome samples, so we assessed the change in proportions of their core bacterial genera (a group of bacteria commonly present in swine and mouse guts [38, 39]). The percentages of confirmed core bacterial genera improved by up to 5.7% after applying cutoffs on query coverage (Table 3). In addition, in the nanopore human data sets, the total percentage of reads classified as *Homo* sp. in the three healthy individual samples were improved by applying cutoffs to query coverage. These results indicated that this strategy may be broadly applicable not only to fungal species, but also to the classification of other eukaryotes and bacteria. One Illumina data set (d1) and one nanopore data set (a5) showed decreased percentages of confirmed taxa after applying query coverage cutoffs, which might be because the core microbiome species are not representing the species identified in the Illumina sample, or due to the low coverage and high error rate of nanopore data.

Downloaded from https://journals.asm.org/journal/mbio on 08 June 2022 by 54.66.212.33.

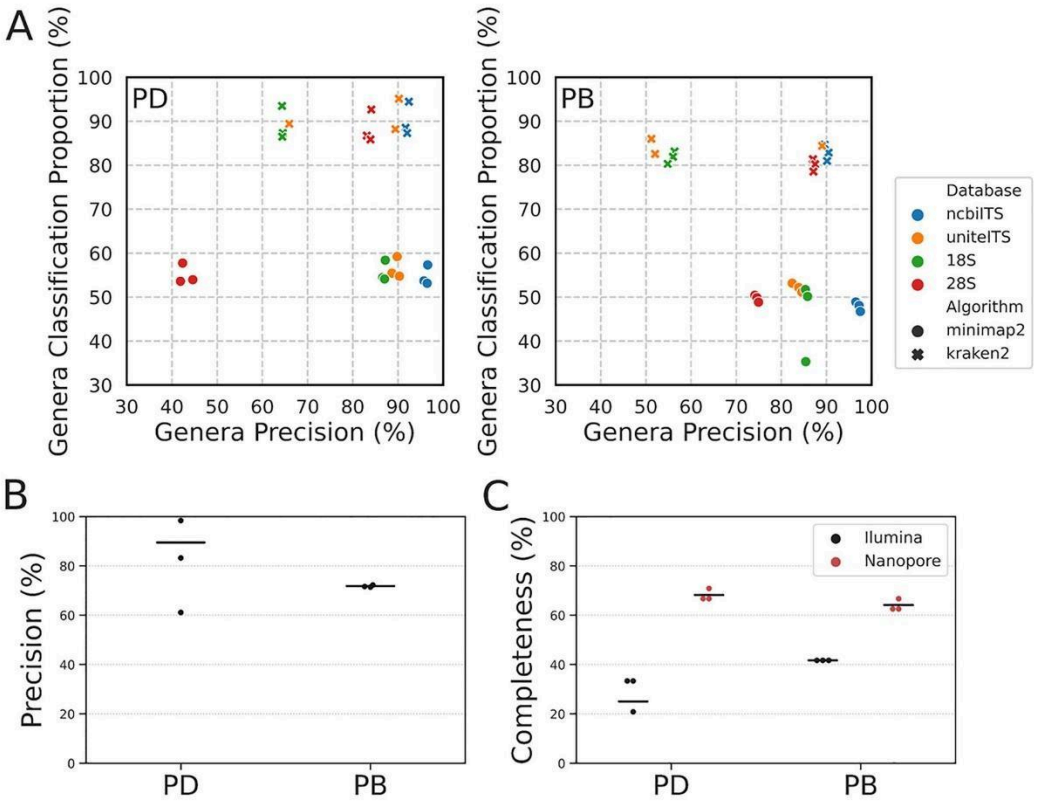


FIG 3 Benchmarking of amplicon data sets. PD = pooled DNA; PB = pooled biomass. (A) Scatterplot representing genus level classification proportion and precision for nanopore amplicon data. (B) Genus level precision of Illumina amplicon data. Classification proportion values for Illumina data were 100% due to the nature of the QIIME2 pipeline (based on the UNITE ITS database). (C) Genus level completeness of both nanopore and Illumina amplicon data sets. The nanopore results are from minimap2 algorithm against the UNITE ITS database.

Benchmarking classification pipelines for amplicon data sets identified advantages of each strategy. We next assessed different strategies for the classification of ITS amplicon data sets. We amplified the ITS region from both mock communities using two different primer pairs and three technical replicates for each sample. Taking advantage of nanopore technology, we performed long-amplicon sequencing of a roughly 3 kb rRNA gene region covering part of the 28S subunit, ITS1, 5.8S subunit, ITS2 and part of the 18S subunit (11). For Illumina sequencing we used the well-established ITS1F-ITS2 amplicon of about 300 bp in length (40). Similar to the analysis of the shotgun data sets, we applied both k-mer and alignment-based approaches to the classification of nanopore amplicon data. We used the pairwise alignment algorithm minimap2 as the alignment algorithm instead of BLAST due to its speed and efficiency. We tested four different databases for classification of long amplicons; the NCBI 18S and 28S databases, and two ITS databases from NCBI and UNITE (30, 41). Overall, we found that the k-mer algorithm returned much higher classification proportion than alignment for each nanopore data set, but the highest precision (~97%) values were achieved by combining the minimap2 alignment algorithm with the NCBI ITS database (Fig. 3A). For Illumina amplicon data sets, we applied the QIIME2 pipeline which is one of the most widely used strategies for ITS classification and community composition

Downloaded from https://journals.asm.org/journal/mbio on 08 June 2022 by 54.66.212.33.

analysis (42). The QIIME2 pipeline groups similar Illumina amplicons into sequence features before classification to reduce the demand on computational resources (43). Since all individual Illumina reads are grouped into sequence features and all the sequence features are classified, the classification proportion of the Illumina amplicon data sets are 100%. We plotted precision rates from the QIIME2 analysis of both the pooled DNA and pooled biomass samples with their means (Fig. 3B). The mean precision values from either Illumina data set were lower than that from k-mer analysis of the respective nanopore data sets.

Although the precision values from amplicon data sets were higher than those from shotgun data sets, the ITS classifications did not identify all genera within the mock community, as shown by our completeness analysis (Fig. 3C). The nanopore amplicons identified 68% (pooled DNA) and 63% (pooled biomass) of the total genera in the mock community, whereas the Illumina amplicon data sets covered only 25% and 41% of the genera, respectively. We suspect that the low completeness from ITS classifications was due in part to the low quality of this particular data set (Table 2) and partially due to biases arising from non-uniform amplification by used primer pairs and different amplicon lengths. However, there were fewer nanopore amplicon reads than in the Illumina amplicon data sets and the completeness from the nanopore data were higher (Fig. 3C). This supports the argument that long amplicons identify a wider range of species and are more accurate in species classification than short amplicons (44, 45).

Cutoffs on query coverage also improve community composition analysis. We next calculated community compositions using the results from BLAST search against Refseq fungal database for all shotgun metagenomics data sets. Community composition refers to the identity and relative abundances of all taxa in a community. Given the observation that use of a restricted database resulted in higher classification precision, we constructed a database containing only the genomes from species within the mock community and aligned all data to the mock community database using BLAST. This forces the precision to 100% as any classification will belong to a species from the mock community. We then BLASTed each data set against this database and calculated the relative abundance of each genus. We defined this as the 'gold standard' for community composition analysis of the mock fungal community (Fig. 4A). We then compared the community composition determined from each combination of algorithms and databases with the gold standard for each data set, and measured their differences using three statistical distance tests: Bhattacharyya distance, relative Euclidean distance and relative entropy (46–48). Consistently, BLASTing sequences against the Refseq fungal database produced community compositions with the highest similarity to the gold standard analysis (Fig. 4B).

To assess whether query coverage cutoffs also improved the community composition analysis of shotgun metagenomics data, we plotted the changes in statistical distance after progressive application of query coverage cutoffs (Fig. 4). After applying cutoffs on the query coverage, the community composition improved in all cases especially for lower cut-off values. The community compositions from pooled DNA Illumina data sets improved and turned out to be the most similar result to the gold standard at query-coverage cutoffs greater than 3–4%, which is consistent with the changes in precision rate shown in Fig. 2D. Overall, our results illustrated that applying cutoffs on query coverage did not only improve the classification accuracy, but also the community composition analysis.

DISCUSSION

Here, we investigated the taxonomic classification of DNA sequences, one of the key steps in all metagenomic workflows, with a particular focus on fungi. After assessing various combinations of algorithms and databases following different sequencing strategies, we found that combining BLAST with the fungal specific Refseq fungal database always resulted in the most precise classifications for all mock fungal community data sets. These classifications were further improved when applying cutoffs on query

Downloaded from https://journals.asm.org/journal/mbio on 08 June 2022 by 54.66.212.33.

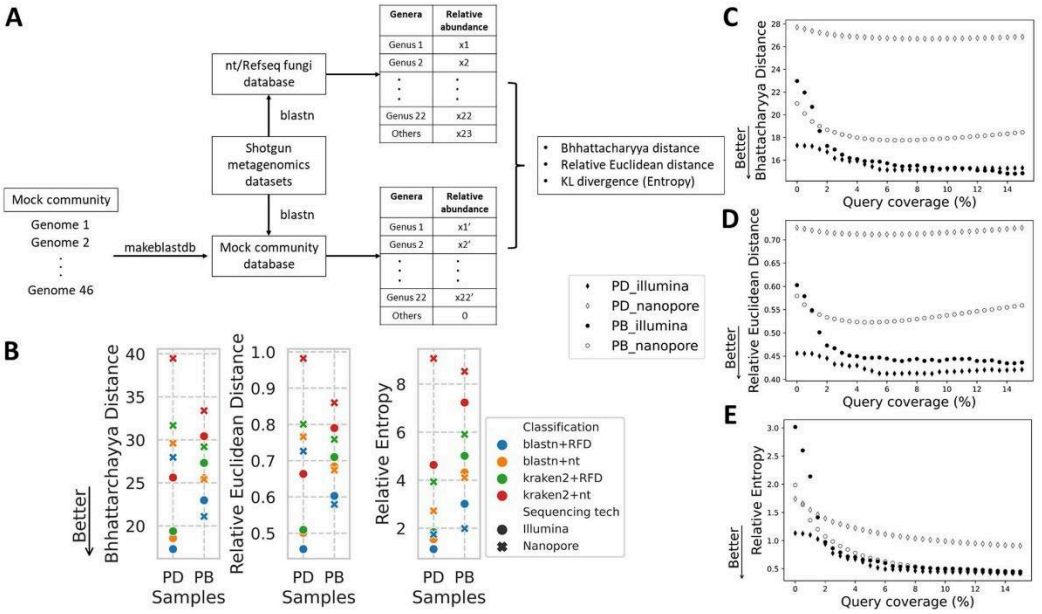


FIG 4 Improving community composition analysis by applying query coverage cutoffs. PD = pooled DNA; PB = pooled biomass; nt = NCBI nucleotide database (29); RFD = RefSeq Fungi database (30) (A) Experimental flowchart for analyzing community compositions. (B) Statistical similarity measures between gold standard community composition and each combination of algorithms and databases. Lower values correspond to greater similarity between the samples and the gold standard. (C) Change in Bhattacharyya distance after applying cutoffs to query coverage for each data set as indicated. The query coverage gap between each dot point is 0.5%. (D) Change in relative Euclidean distance after applying cutoffs to query coverage for each data set. The gap between each dot point is 0.5%. (E) Change in relative entropy after applying cutoffs on query coverage for each data set. The gap between each dot point is 0.5%.

coverage, including positive flow on effects on downstream community composition analysis from shotgun metagenomics data sets.

Although sampling strategies and DNA extraction substantially influence the outcome of species classifications (49–51), choosing an appropriate sequencing strategy is a major step toward accurately profiling a sample. Overall, our results suggested that both short- and long-read shotgun metagenomics data sets have comparable accuracy, and both achieved higher accuracy than amplicon data sets. However, Illumina shotgun data sets require additional steps to assemble reads into contigs before querying them against a database, and to map all reads back to the assembly to quantify the coverage. These processes are necessary to achieve accurate classification from longer contigs (52), but take longer than analysis of long-read shotgun data. For amplicon data, long range amplicons performed better than short ITS data in classification accuracy and completeness (Fig. 3), consistent with other studies (44, 45). The overall completeness from analyzing amplicon data sets is much lower than from shotgun data sets (Fig. 3C). This may be because we used far less amplicon data for benchmarking classification pipelines, and an incomplete database which does not contain all taxa present in our mock community. On the other hand, amplification biases are another major source of poor community recovery. These biases come mainly from inconsistent amplification of the barcoding regions of different species, caused by copy number variations and different primer binding specificities (53, 54). Also, particularly for fungi, variations in barcode (amplicon) length are the major source of bias in recording fungal community compositions, with longer barcodes being underrepresented (55). Overall, analysis of the long-read shotgun data sets returned the most accurate fungal classification.

Downloaded from https://journals.asm.org/journal/mbio on 08 June 2022 by 54.66.212.33.

Next, we found that the alignment algorithm (BLAST) outperformed the k-mer based approach (kraken2) in classification accuracy (24, 56). We also compared the effects of progressive cutoffs to major alignment parameters for classification of shotgun metagenomics data. We found that applying read length or read quality cutoffs did not improve the precision of the classification for all shotgun data sets (Fig. 2A, Fig. S1). This observation contrasts with a previous study based on simulated data which claimed that the use of long reads improves the accuracy of classification (57). Cutoffs on pident slightly improved the classification accuracy for Illumina data sets, but the error-prone nature of the nanopore data (~10% error rate with Guppy 4 basecalling) affected the results evident as a breakdown of precision when pident cutoffs reached 90% (Fig. 2C).

We found that query coverage cutoffs that filtered out ~20% of BLAST results worked most effectively for improving the classification (Fig. 2D, Table 3). Unlike E value which weights gaps and mismatches as the major factors affecting alignment quality, query coverage weights the query length as well as the number of identical matches in assessing alignment quality. In this case, we can eliminate spurious alignments that are due to a small proportion of reads with high fidelity to the reference which are commonly present in metagenomics data sets. Interestingly, higher cutoffs on query coverage (10–20%) are observed from mock community data sets than real environmental data sets given the same 20% filtering threshold on BLAST results, including few extremely low thresholds of query coverage in the contigs from Illumina shotgun data sets. Other studies that used simulated data to generate metagenomics contigs for classification employed 90% query coverage cutoffs (58–60). This clearly indicates an important difference between real mock community data sets and simulated data based on reference databases, because contigs from simulated reads are artificially more similar to the reference genomes than the real data. Therefore, the use of simulated data may overestimate the specificity of the alignment, which leads to overoptimistic estimation of the classification. Together with the different results from read length and read quality cutoffs, these observations highlight the differences between the use of real environmental data and simulated data in benchmarking studies, especially for the classification of complex microbial communities.

There were some differences in results after analyzing pooled DNA and pooled biomass samples after comparing statistical distance with the gold standard. pooled DNA samples pooled DNA of individual species together while pooled biomass samples pooled equal amount of fungal tissue before the DNA extraction. After applying cutoffs on query coverage, both Bhattacharyya distance and Euclidean distance between the best practice (BLAST + Refseq fungal database) and the gold standard classification only showed marginal decrease in pooled DNA samples, and slowly reversed as the cutoffs increase (Fig. 4). In contrast, these distances from pooled biomass samples decreased more significantly and never reversed, indicating a clear improvement of the community composition analysis (Fig. 4). This may be because about 1/3 of reads were classified as *Candida* in the pooled DNA sample while only 10% reads were classified as *Candida* in pooled biomass sample. The difference in the relative abundance of this single genus between the community compositions of the gold standard and the best practice is much higher than that of other low abundance genera, and is thus more influential in calculation of the overall statistical distance of each data set.

Following the effectiveness of the query coverage, the next question is how to bring low quality but high coverage alignments into consideration? The winner-takes-all selection strategy could itself be redesigned, as highly conserved genome regions from closely related species often resulted in very similar alignment scores between the best alignment and other top alignments of each query. In this case, a weighting statistic and the relative probabilities of multiple top taxonomic assignments can be explored to replace the best-hit-takes-all strategy. This will be particularly useful in conjunction with the rapid expansion of fungal genome databases.

Next to the use of the best classification tool, choosing the most appropriate database significantly influences analysis outcomes (25, 26). Based on our observations, we

Downloaded from https://journals.asm.org/journal/mbio on 08 June 2022 by 54.66.212.33.

suggest that 'prior knowledge' of the data set should guide the choice of database as this will improve the accuracy of taxonomic classifications. For example, our results suggested that the restricted database resulted in more accurate fungal classifications for shotgun metagenomics data sets. This strategy might be appropriate if queries are initially binned into kingdoms before a more in-depth analysis using kingdom specific databases. Also, Kaehler et al. (61) incorporated environment-specific taxonomic abundance information into the analysis of amplicon data sets and showed that these improve classification accuracy. Similar approaches can be applied to metagenomic data sets. In addition, machine learning strategies are becoming increasingly popular for analyzing genomic data. Here, taxonomic classifiers could be trained on classified sequence data sets before being applied to communities with similar compositions to the training data sets, or to identify target species from complex communities (56, 62).

MATERIALS AND METHODS

Code availability. All detailed commands and scripts used in each step are summarized in https://github.com/Yiheng323/Inferring-species-compositions-of-complex-fungal-communities-from-long-and-short-read-sequence-data.

Fungal harvesting, DNA extraction and construction of mock communities. Selected fungal strains were grown for 48 h at 27°C on Sabouraud dextrose agar. For the species in the pooled DNA community, an inoculation loop of fungal cells was scraped into a 1.5 mL microcentrifuge tube and ground with a pestle in liquid nitrogen. Genomic DNA was extracted using the Zymo Research *Quick*-DNA Fungal/Bacterial Miniprep kit (cat. no. D6005 Zymo Research, Irvine, CA, USA). First, BashingBead Buffer was added to the ground fungal cells and vortexed. The mixture was filtered through a Zymo-SpinTM III-F Column and the filtrate was combined with Genomic Lysis Buffer. The mixture was filtered through a Zymo-Spin IICR Column and washed with DNA Pre-Wash buffer and g-DNA Wash Buffer. The DNA was eluted in nuclease-free water. DNA concentrations were measured using the DeNovix dsDNA Broad Range kit (DeNovix, Wilmington, DE, USA) and 250 ng of DNA from each strain was then pooled to create the final community.

For the pooled biomass community, two inoculation loops of fungal cells of each species in the mock community were scraped into a ceramic mortar. Liquid nitrogen was poured into the mortar and the fungal mixture was ground into a fine powder. DNA was extracted using the Qiagen DNeasy PowerMax Soil kit (cat. no. 12988-10 Qiagen, Hilden, Germany). PowerBead Solution and Solution C1 were added to the ground fungal material, vortexed and centrifuged. The supernatant was then added to Solution C2, mixed and centrifuged, which was then repeated with Solution C3. The resulting supernatant was combined with Solution C4 and centrifuged through a column. The column was then washed twice with Solution C5. Final DNAs were eluted in nuclease free water and the concentration measured using the DeNovix dsDNA Broad Range kit.

Library preparation and sequencing. The ITS1 regions of the rRNA gene were amplified with the universal fungal primers, ITS1F (CTTGGTCATTTAGAGGAAGTAA) and ITS2 (GCTGCGTTCTTCATCGATGC) (40). Sequencing of PCR amplicons was conducted on the MiSeq System (Illumina, San Diego, CA, USA) at the Australian Genome Research Facility. The Illumina bcl2fastq 2.18.0.12 pipeline was used to generate the sequence data. Paired-end reads 2 × 300 bp were generated up to 0.15 GB per sample for amplicon data. The Illumina amplicon data were then imported directly into QIIME2 for analysis. For shotgun Illumina data sets, we employed the same sequencing pipeline as the amplicon data, with MiSeq and the bcl2fastq 2.18.0.12 pipeline for the 2 × 300 bp paired-end reads. Raw shotgun Illumina reads were trimmed of adapters with Trimomatic (63). Quality controlled, paired-end reads were merged and assembled into metagenomics contigs using IDBA_UD (64), which is suitable for data sets with uneven sequencing depths of each species. After assembly, raw reads were mapped back to the contigs using bwa-mem (65), and the bam files were generated and sorted from sam files using samtools (66). Bedtools (67) was used for generating coverage for each contig, and we used python numpy and pandas module to calculate the average coverage for each contig.

For Nanopore sequencing (both shotgun and amplicon sequencing), we used Ligation Sequencing 1D SQK-LSK108 and Native Barcoding Expansion (PCR-free) EXP-NBD103 kits from ONT (UK), as adapted by Hu and Schwessinger (68), which was modified from the manufacturer's instructions with the omission of the DNA fragmentation and DNA repair steps. DNA was first cleaned up using 1× volume of Agencourt AMPure XP beads (cat. no. A63881, Beckman Coulter, Indianapolis, IN, USA) following the manufacturer's instructions. We then eluted the DNA from the beads in 51 μL nuclease free water and quantified it using NanoDrop and Quibit Fluorometers (Thermo Fisher Scientific, Waltham, MA, USA). DNA was end-repaired (NEBNext Ultra II End-Repair/dA-tailing Module, cat. No. E7546), cleaned up with1x volume beads (AMPure XP beads), and eluted in 31 μL nuclease free water. The barcoding reaction was performed by adding 2 μL of each native barcode and 20 μL NEB Blunt/TA Master Mix (cat. No. M0367) to 18 μL DNA, mixing gently and incubating at room temperature for 10 min. A 1× volume (40 μL) Agencourt AMPure XP beads cleanup was then performed, and the DNA was eluted in 15 μL nuclease free water. Ligation was performed by adding 20 μL Barcode Adapter Mix (EXP-NBD103 Native Barcoding Expansion kit, ONT, UK), 20 μL NEBNext Quick Ligation Reaction Buffer, and Quick T4 DNA Ligase (cat. No. E6056) to the 50 μL pooled equimolar barcoded DNA, mixing gently and incubating at

Downloaded from https://journals.asm.org/journal/mbio on 08 June 2022 by 54.66.212.33.

room temperature for 10 min. The adapter-ligated DNA was cleaned-up by adding 0.4× volume (40 μL) of Agencourt AMPure XP beads, incubating for 5 min at room temperature and resuspending the pellet twice in 140 μL ABB provided in the SQK-LSK108 kit. The purified-ligated DNA was resuspended by adding 15 μL ELB provided in the SQK-LSK108 kit and resuspending the beads. The beads were pelleted again, and the supernatant sequencing library was transferred to a new 0.5 mL DNA LoBind tube (Eppendorf, Germany). Nanopore sequencing was carried out on a MinION MK1b device using R9.4.1 Flowcells. Raw fast5 files were barcode demultiplexed by deepbiner (ONT), then basecalled by Guppy (v3.6.0, ONT, UK). Quality passed reads in fastq files were trimmed of adapters and barcodes using qcat (ONT, UK). For the long amplicon data, we filtered out reads less than 2000 bp. All sequencing data were submitted to the NCBI Sequence Read Archive (SRA) under the Bioproject PRJNA725368 including eight accessions: SRX10705648, SRX10705649, SRX10705650, SRX10705651, SRX10705695, SRX10705696, SRX10705697 and SRX10705698.

Genome assembly. While generating the reference genome database we found that there were no reference genomes for *Diutina rugosa* (former *Candida rugosa*), *Diutina mesorugosa* (former *Candida mesorugosa*), and *Filobasidium magnus* (former *Cryptococcus magnus*), so we performed nanopore sequencing on pure DNA from each species and assembled their draft genomes. These assemblies were of sufficient contiguity and quality (Table S1), so we added the new draft genomes into the reference database.

The nanopore data for *Diutina rugosa*, *Diutina mesorugosa,* and *Filobasidium magnus* were generated individually using the Ligation Sequencing 1D SQK-LSK108 kit alone, and from independent flowcells. Data from each flowcell was basecalled and quality filtered using the same pipeline as described above. We got roughly 40× genome coverage for *Diutina rugosa* and *Diutina mesorugosa*, and 20× coverage for *Filobasidium magnus*. Draft genomes were assembled with Flye (69) using default parameters and an estimated genome size of 20 Mb. After assembly, the contigs were polished 10 times with Racon (70) using nanopore reads, followed by a single polishing step with Medaka (ONT). The polished assemblies were assessed for completeness using BUSCO (71). The assembly statistics were reported from Flye.

Database constructions. For shotgun metagenomics analysis, we used three BLAST database and three kraken databases. Two databases (NCBI nucleotide database and Refseq fungal database) are from the same NCBI source, downloaded in May 2019. BLAST and kraken2 nt databases were downloaded using the updateblastdb.pl script from BLAST+ package (72) and the kraken2-build command (21), respectively. The fasta files of the RefSeq fungal database were downloaded from NCBI and converted to a BLAST database using the makblastdb command from the BLAST+ package (72), and were added to the kraken2 database library using the kraken2-build command.

To generate the mock community database containing only those species in the mock community for BLAST, we downloaded the genomes of all mock community species from NCBI according to their accessions (Table 1) and concatenated them with the three newly assembled genomes of *Diutina rugosa*, *Diutina mesorugosa,* and *Filobasidium magnus*. Following the previous pipeline (73), we then performed a kraken2 search to identify the potential contaminated regions in the concatenated fasta and masked those regions using bedtools (67). This kraken2 search used the standard kraken2 database which was build using the kraken2-build command. We also masked the low complexity regions using dustmasker from the BLAST+ package (72). To enable new genomes to be indexed by blastn, we updated the taxonomic map file by adding the fasta headers of the three new genomes and manually assigned their taxonomic IDs in the file. Lastly, we used the makeblastdb program to construct the mock community database.

For amplicon data analysis, we used two versions of the fungal ITS database from NCBI and UNITE, plus the fungal 18S, 28S database from NCBI. All of these were downloaded in fasta format in February 2020 and added to the kraken2 database library using the kraken2-build command.

Data analysis. For shotgun metagenomics data sets, we first used blastn (version 2.10.1) and kraken2 (version 2.0.8) to assign the NCBI taxonomic ID for each Illumina contig or Nanopore read. During classification, we found that a potential contaminant species *Purpureocillium lilacinum* was present in all samples in significant abundance (10–20%). Therefore, we added this species to the species list. The best hit from BLAST, or the species with the highest k-mer counts for each read and/or contig, was retained for further analysis. After classification, we used the python pandas module to merge information from different output files, and used the ete3 module (74) to assign taxonomic information to each read or contig. The relative abundances of each classification were calculated based on the total length of nanopore reads or total coverage of Illumina contigs. We used the python numpy and math modules for all statistical analysis.

For amplicon data sets, we sequenced each sample with three technical replicates. The classification workflow was different for data sets with different sequencing technologies. We used only the QIME2 workflow plus the UNITE database for the Illumina amplicon data, since it is the widely used method for classification. The paired-end reads were denoised using the DADA2 (75) plugin and assigned taxonomic information using the q2-feature-classifier (76) plugin. The QIME2 classifier was trained by the database sequence before classification. The classification output .qzv files were visualized by the QIME2 view website (https://view.qiime2.org/) and the feature-frequency csv file was extracted from the website. We then used the python numpy and math modules for the mathematical analysis and used the seaborn module to generate figures.

For nanopore amplicon data sets, we used kraken2 as the k-mer based algorithm and minimap2 as the alignment-based algorithm. The kraken2 command is the same as the kraken2 analysis for the shotgun metagenomics data sets but uses different databases. For the minimap2 analysis, we extracted the accessions of the best hits from the output files, and searched their corresponding taxonomic IDs from

Downloaded from https://journals.asm.org/journal/mbio on 08 June 2022 by 54.66.212.33.

the NCBI taxonomic map (downloaded from https://ftp.ncbi.nih.gov/pub/taxonomy/accession2taxid/nucl_wgs.accession2taxid.gz, in June 2020) using the python pandas module. We then merged information from different output files and used the ete3 module to assign taxonomic information to each read.

Availability of data and material. All sequencing data were submitted to NCBI Short Read Archive (SRA) under the BioProject PRJNA725368 including eight accessions: SRX10705648, SRX10705649, SRX10705650, SRX10705651, SRX10705695, SRX10705696, SRX10705697 and SRX10705698.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, JPG file, 0.6 MB.

TABLE S1, PPTX file, 0.04 MB.

ACKNOWLEDGMENTS

We thank Dr. Yu Lin for the insightful discussion and valuable suggestions to our manuscript, and the editor and reviewer for excellent feedback that significantly improved the manuscript. This work was supported by computational resources provided by the Australian Government through the National Computational Infrastructure (NCI) under the ANU Merit Allocation Scheme.

We have no conflicts of interest to declare.

This study was supported by a National Health and Medical Research Council of Australia (NH&MRC) grant [number GNT1121936] to W.M., B.S. is supported by an Australian Research Council Future Fellowship FT180100024, and Y.H., E.S., J.R., and B.S. are supported by The Hermon Slade Foundation grant HSF_17_04.

WM, ES, BS, and JPR conceived the study and designed experiments. YH, LI, and WTVH prepared the samples and generated sequencing data. YH, TE, and AG performed the bioinformatics analysis. ES provided feedback on statistical analysis. All authors contributed to data analysis and manuscript writing. All authors read and approved the final manuscript.

REFERENCE

1. Hawksworth DL, Lücking R. 2017. Fungal diversity revisited: 2.2 to 3.8 million species. Microbiol Spectr 5:10. https://doi.org/10.1128/microbiolspec.FUNK-0052-2016.
2. Cheek M, Nic Lughadha E, Kirk P, Lindon H, Carretero J, Looney B, Douglas B, Haelewaters D, Gaya E, Llewellyn T, Ainsworth AM, Gafforov Y, Hyde K, Crous P, Hughes M, Walker BE, Campostrini Forzza R, Wong KM, Niskanen T. 2020. New scientific discoveries: plants and fungi. Plants People Planet 2:371–388. https://doi.org/10.1002/ppp3.10148.
3. Fisher MC, Gurr SJ, Cuomo CA, Blehert DS, Jin H, Stukenbrock EH, Stajich JE, Kahmann R, Boone C, Denning DW, Gow NAR, Klein BS, Kronstad JW, Sheppard DC, Taylor JW, Wright GD, Heitman J, Casadevall A, Cowen LE. 2020. Threats posed by the fungal kingdom to humans, wildlife, and agriculture. mBio 11:e00449-20. https://doi.org/10.1128/mBio.00449-20.
4. Scheele BC, Pasmans F, Skerratt LF, Berger L, Martel A, Beukema W, Acevedo AA, Burrowes PA, Carvalho T, Catenazzi A, la Riva ID, Fisher MC, Flechas SV, Foster CN, Frías-Álvarez P, Garner TWJ, Gratwicke B, Guayasamin JM, Hirschfeld M, Kolby JE, Kosch TA, Marca EL, Lindenmayer DB, Lips KR, Longo AV, Maneyro R, McDonald CA, Mendelson J, Palacios-Rodriguez P, Parra-Olea G, Richards-Zawacki CL, Rödel M-O, Rovito SM, Soto-Azat C, Toledo LF, Voyles J, Weldon C, Whitfield SM, Wilkinson M, Zamudio KR, Canessa S. 2019. Amphibian fungal panzootic causes catastrophic and ongoing loss of biodiversity. Science 363:1459–1463. https://doi.org/10.1126/science.aav0379.
5. Naranjo-Ortiz MA, Gabaldón T. 2019. Fungal evolution: diversity, taxonomy and phylogeny of the Fungi. Biol Rev Camb Philos Soc 94:2101–2137. https://doi.org/10.1111/brv.12550.
6. Valverde ME, Hernández-Pérez T, Paredes-López O. 2015. Edible mushrooms: improving human health and promoting quality life. Int J Microbiol 2015:376387. https://doi.org/10.1155/2015/376387.
7. Kong HH, Segre JA. 2020. Cultivating fungal research. Science 368:365–366. https://doi.org/10.1126/science.aaz8086.
8. Jones MDM, Forn I, Gadelha C, Egan MJ, Bass D, Massana R, Richards TA. 2011. Discovery of novel intermediate forms redefines the fungal tree of life. Nature 474:200–203. https://doi.org/10.1038/nature09984.
9. Adl SM, Simpson AGB, Lane CE, Lukeš J, Bass D, Bowser SS, Brown MW, Burki F, Dunthorn M, Hampl V, Heiss A, Hoppenrath M, Lara E, Le Gall L, Lynn DH, McManus H, Mitchell EAD, Mozley-Stanridge SE, Parfrey LW, Pawlowski J, Rueckert S, Shadwick L, Shadwick L, Schoch CL, Smirnov A, Spiegel FW. 2012. The revised classification of eukaryotes. J Eukaryot Microbiol 59:429–514. https://doi.org/10.1111/j.1550-7408.2012.00644.x.
10. Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, Chen W, Consortium FB, List FBCA, Bolchacova E, Voigt K, Crous PW, Miller AN, Wingfield MJ, Aime MC, An K-D, Bai F-Y, Barreto RW, Begerow D, Bergeron M-J, Blackwell M, Boekhout T, Bogale M, Boonyuen N, Burgaz AR, Buyck B, Cai L, Cai Q, Cardinali G, Chaverri P, Coppins BJ, Crespo A, Cubas P, Cummings C, Damm U, de Beer ZW, de Hoog GS, Del-Prado R, Dentinger B, Diéguez-Uribeondo J, Divakar PK, Douglas B, Dueñas M, Duong TA, Eberhardt U, Edwards JE, Elshahed MS, Fliegerova K, Furtado M, García MA, Fungal Barcoding Consortium Author List, et al. 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proc Natl Acad Sci U S A 109:6241–6246. https://doi.org/10.1073/pnas.1117018109.
11. Raja HA, Miller AN, Pearce CJ, Oberlies NH. 2017. fungal identification using molecular tools: a primer for the natural products research community. J Nat Prod 80:756–770. https://doi.org/10.1021/acs.jnatprod.6b01085.
12. Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. 2015. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res 43:e37. https://doi.org/10.1093/nar/gku1341.
13. Frau A, Kenny JG, Lenzi L, Campbell BJ, Ijaz UZ, Duckworth CA, Burkitt MD, Hall N, Anson J, Darby AC, Probert CSJ. 2019. DNA extraction and amplicon production strategies deeply inf luence the outcome of gut mycobiome studies. Sci Rep 9:9328. https://doi.org/10.1038/s41598-019-44974-x.
14. De Filippis F, Laiola M, Blaiotta G, Ercolini D. 2017. Different amplicon targets for sequencing-based studies of fungal diversity. Appl Environ Microbiol 83:e00905-17. https://doi.org/10.1128/AEM.00905-17.

Downloaded from https://journals.asm.org/journal/mbio on 08 June 2022 by 54.66.212.33.

15. Heeger F, Bourne EC, Baschien C, Yurkov A, Bunk B, Spröer C, Overmann J, Mazzoni CJ, Monaghan MT. 2018. Long-read DNA metabarcoding of ribosomal RNA in the analysis of fungi from aquatic environments. Mol Ecol Resour 18:1500–1514. https://doi.org/10.1111/1755-0998.12937.
16. D'Andreano S, Cuscó A, Francino O. 2021. Rapid and real-time identification of fungi up to species level with long amplicon nanopore sequencing from clinical samples. Biol Methods Protoc 6:bpaa026. https://doi.org/10.1093/biomethods/bpaa026.
17. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. 2017. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol 35:833–844. https://doi.org/10.1038/nbt.3935.
18. Marcelino VR, Clausen PTLC, Buchmann JP, Wille M, Iredell JR, Meyer W, Lund O, Sorrell TC, Holmes EC. 2020. CCMetagen: comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data. Genome Biol 21:103. https://doi.org/10.1186/s13059-020-02014-2.
19. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.
20. Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15:R46. https://doi.org/10.1186/gb-2014-15-3-r46.
21. Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. Genome Biol 20:257. https://doi.org/10.1186/s13059-019-1891-0.
22. Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, Maharjan S, Mailyan A, Manghi P, Scholz M, Thomas AM, Valles-Colomer M, Weingart G, Zhang Y, Zolfo M, Huttenhower C, Franzosa EA, Segata N. 2021. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. Elife 10:e65088. https://doi.org/10.7554/eLife.65088.
23. Zielezinski A, Girgis HZ, Bernard G, Leimeister C-A, Tang K, Dencker T, Lau AK, Röhling S, Choi JJ, Waterman MS, Comin M, Kim S-H, Vinga S, Almeida JS, Chan CX, James BT, Sun F, Morgenstern B, Karlowski WM. 2019. Benchmarking of alignment-free sequence comparison methods. Genome Biol 20:144. https://doi.org/10.1186/s13059-019-1755-7.
24. McIntyre ABR, Ounit R, Afshinnekoo E, Prill RJ, Hénaff E, Alexander N, Minot SS, Danko D, Foox J, Ahsanuddin S, Tighe S, Hasan NA, Subramanian P, Moffat K, Levy S, Lonardi S, Greenfield N, Colwell RR, Rosen GL, Mason CE. 2017. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. Genome Biol 18:182. https://doi.org/10.1186/s13059-017-1299-7.
25. Ye SH, Siddle KJ, Park DJ, Sabeti PC. 2019. Benchmarking Metagenomics Tools for Taxonomic Classification. Cell 178:779–794. https://doi.org/10.1016/j.cell.2019.07.010.
26. Nasko DJ, Koren S, Phillippy AM, Treangen TJ. 2018. RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. Genome Biol 19:165. https://doi.org/10.1186/s13059-018-1554-6.
27. R Marcelino V, Holmes EC, Sorrell TC. 2020. The use of taxon-specific reference databases compromises metagenomic classification. BMC Genomics 21:184. https://doi.org/10.1186/s12864-020-6592-2.
28. Heeger F, Wurzbacher C, Bourne EC, Mazzoni CJ, Monaghan MT. 2019. Combining the 5.8S and ITS2 to improve classification of fungi. Methods Ecol Evol 10:1702–1711. https://doi.org/10.1111/2041-210X.13266.
29. NCBI Resource Coordinators. 2018. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 46:D8–D13. https://doi.org/10.1093/nar/gkx1095.
30. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733–D745. https://doi.org/10.1093/nar/gkv1189.
31. Reynolds NK, Jusino MA, Stajich JE, Smith ME. 2021. Understudied, underrepresented, and unknown: methodological biases that limit detection of early diverging fungi from environmental samples. Mol Ecol Resour https://doi.org/10.1111/1755-0998.13540.
32. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402. https://doi.org/10.1093/nar/25.17.3389.
33. Xiao L, Feng Q, Liang S, Sonne SB, Xia Z, Qiu X, Li X, Long H, Zhang J, Zhang D, Liu C, Fang Z, Chou J, Glanville J, Hao Q, Kotowska D, Colding C, Licht TR, Wu D, Yu J, Sung JJY, Liang Q, Li J, Jia H, Lan Z, Tremaroli V, Dworzynski P, Nielsen HB, Bäckhed F, Doré J, Le Chatelier E, Ehrlich SD, Lin JC, Arumugam M, Wang J, Madsen L, Kristiansen K. 2015. A catalog of the mouse gut metagenome. Nat Biotechnol 33:1103–1108. https://doi.org/10.1038/nbt.3353.
34. Hu Y, Green GS, Milgate AW, Stone EA, Rathjen JP, Schwessinger B. 2019. Pathogen Detection and Microbiome Analysis of Infected Wheat Using a Portable DNA Sequencer. Phytobiomes J 3:92–101. https://doi.org/10.1094/PBIOMES-01-19-0004-R.
35. Irinyi L, Hu Y, Hoang MTV, Pasic L, Halliday C, Jayawardena M, Basu I, McKinney W, Morris AJ, Rathjen J, Stone E, Chen S, Sorrell TC, Schwessinger B, Meyer W. 2020. Long-read sequencing based clinical metagenomics for the detection and confirmation of Pneumocystis jirovecii directly from clinical specimens: a paradigm shift in mycological diagnostics. Med Mycol 58:650–660. https://doi.org/10.1093/mmy/myz109.
36. Xiao L, Estellé J, Kiilerich P, Ramayo-Caldas Y, Xia Z, Feng Q, Liang S, Pedersen AØ, Kjeldsen NJ, Liu C, Maguin E, Doré J, Pons N, Le Chatelier E, Prifti E, Li J, Jia H, Liu X, Xu X, Ehrlich SD, Madsen L, Kristiansen K, Rogel-Gaillard C, Wang J. 2016. A reference gene catalogue of the pig gut microbiome. Nat Microbiol 1:1–6. https://doi.org/10.1038/nmicrobiol.2016.161.
37. Donovan PD, Gonzalez G, Higgins DG, Butler G, Ito K. 2018. Identification of fungi in shotgun metagenomics datasets. PLoS One 13:e0192898. https://doi.org/10.1371/journal.pone.0192898.
38. Holman DB, Brunelle BW, Trachsel J, Allen HK. 2017. Meta-analysis To Define a Core Microbiota in the Swine Gut. mSystems 2:e00004-17. https://doi.org/10.1128/mSystems.00004-17.
39. Wang J, Lang T, Shen J, Dai J, Tian L, Wang X. 2019. Core gut bacteria analysis of healthy mice. Front Microbiol 10:887. https://doi.org/10.3389/fmicb.2019.00887.
40. White T, Bruns T, Lee S, Taylor J, Innis M, Gelfand D, Sninsky J. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In PCR Protocols: a Guide to Methods and Applications 315–322.
41. Nilsson RH, Larsson K-H, Taylor AFS, Bengtsson-Palme J, Jeppesen TS, Schigel D, Kennedy P, Picard K, Glöckner FO, Tedersoo L, Saar I, Köljalg U, Abarenkov K. 2019. The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. Nucleic Acids Res 47:D259–D264. https://doi.org/10.1093/nar/gky1022.
42. Bharti R, Grimm DG. 2021. Current challenges and best-practice protocols for microbiome analysis. Brief Bioinform 22:178–193. https://doi.org/10.1093/bib/bbz155.
43. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, et al. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37:852–857. https://doi.org/10.1038/s41587-019-0209-9.
44. Johnson JS, Spakowicz DJ, Hong B-Y, Petersen LM, Demkowicz P, Chen L, Leopold SR, Hanson BM, Agresta HO, Gerstein M, Sodergren E, Weinstock GM. 2019. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun 10:5029. https://doi.org/10.1038/s41467-019-13036-1.
45. Krehenwinkel H, Pomerantz A, Henderson JB, Kennedy SR, Lim JY, Swamy V, Shoobridge JD, Graham N, Patel NH, Gillespie RG, Prost S. 2019. Nanopore sequencing of long ribosomal DNA amplicons enables portable and simple biodiversity assessments with high phylogenetic resolution across broad taxonomic scale. GigaScience 8:giz006. https://doi.org/10.1093/gigascience/giz006.
46. Dokmanic I, Parhizkar R, Ranieri J, Vetterli M. 2015. Euclidean distance matrices: essential theory, algorithms, and applications. IEEE Signal Process Mag 32:12–30. https://doi.org/10.1109/MSP.2015.2398954.
47. Aherne FJ, Thacker NA, Rockett PI. 1998. The Bhattacharyya metric as an absolute similarity measure for frequency coded data. Kybernetika 34:363–368.

Downloaded from https://journals.asm.org/journal/mbio on 08 June 2022 by 54.66.212.33.

48. MacKay DJC. 2003. Information Theory, Inference and Learning Algorithms. Cambridge University Press.
49. Henderson G, Cox F, Kittelmann S, Miri VH, Zethof M, Noel SJ, Waghorn GC, Janssen PH. 2013. Effect of DNA extraction methods and sampling techniques on the apparent structure of cow and sheep rumen microbial communities. PLoS One 8:e74787. https://doi.org/10.1371/journal.pone.0074787.
50. Davis A, Kohler C, Alsallaq R, Hayden R, Maron G, Margolis E. 2019. Improved yield and accuracy for DNA extraction in microbiome studies with variation in microbial biomass. Biotechniques 66:285–289. https://doi.org/10.2144/btn-2019-0016.
51. Douglas CA, Ivey KL, Papanicolas LE, Best KP, Muhlhausler BS, Rogers GB. 2020. DNA extraction approaches substantially influence the assessment of the human breast milk microbiome. Sci Rep 10:123. https://doi.org/10.1038/s41598-019-55568-y.
52. Tamames J, Cobo-Simón M, Puente-Sánchez F. 2019. Assessing the performance of different approaches for functional and taxonomic annotation of metagenomes. BMC Genomics 20:960. https://doi.org/10.1186/s12864-019-6289-6.
53. Silverman JD, Bloom RJ, Jiang S, Durand HK, Dallow E, Mukherjee S, David LA. 2021. Measuring and mitigating PCR bias in microbiota datasets. PLoS Comput Biol 17:e1009113. https://doi.org/10.1371/journal.pcbi.1009113.
54. Nearing JT, Comeau AM, Langille MGI. 2021. Identifying biases and their potential solutions in human microbiome studies. Microbiome 9:113. https://doi.org/10.1186/s40168-021-01059-0.
55. Castaño C, Berlin A, Brandström Durling M, Ihrmark K, Lindahl BD, Stenlid J, Clemmensen KE, Olson Å. 2020. Optimized metabarcoding with Pacific biosciences enables semi-quantitative analysis of fungal communities. New Phytol 228:1149–1158.
56. Liang Q, Bible PW, Liu Y, Zou B, Wei L. 2020. DeepMicrobes: taxonomic classification for metagenomics with deep learning. NAR Genomics Bioinforma 2:lqaa009. https://doi.org/10.1093/nargab/lqaa009.
57. Pearman WS, Freed NE, Silander OK. 2020. Testing the advantages and disadvantages of short- and long- read eukaryotic metagenomics using simulated reads. BMC Bioinformatics 21:220. https://doi.org/10.1186/s12859-020-3528-4.
58. Mallawaarachchi V, Wickramarachchi A, Lin Y. 2020. GraphBin: refined binning of metagenomic contigs using assembly graphs. Bioinformatics 36:3307–3313. https://doi.org/10.1093/bioinformatics/btaa180.
59. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. 2014. Binning metagenomic contigs by coverage and composition. Nat Methods 11:1144–1146. https://doi.org/10.1038/nmeth.3103.
60. Wickramarachchi A, Mallawaarachchi V, Rajan V, Lin Y. 2020. MetaBCC-LR: metagenomics binning by coverage and composition for long reads. Bioinformatics 36:i3–i11. https://doi.org/10.1093/bioinformatics/btaa441.
61. Kaehler BD, Bokulich NA, McDonald D, Knight R, Caporaso JG, Huttley GA. 2019. Species abundance information improves sequence taxonomy classification accuracy. Nat Commun 10:4643. https://doi.org/10.1038/s41467-019-12669-6.
62. Bokulich NA, Dillon MR, Bolyen E, Kaehler BD, Huttley GA, Caporaso JG. 2018. q2-sample-classifier: machine-learning tools for microbiome classification and regression. JOSS 3:934. https://doi.org/10.21105/joss.00934.
63. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170.
64. Peng Y, Leung HCM, Yiu SM, Chin FYL. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28:1420–1428. https://doi.org/10.1093/bioinformatics/bts174.
65. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760. https://doi.org/10.1093/bioinformatics/btp324.
66. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352.
67. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. https://doi.org/10.1093/bioinformatics/btq033.
68. Hu Y, Schwessinger B. 2018. Amplicon sequencing using MinION optimized from 1D native barcoding genomic DNA. protocols.io. https://www.protocols.io/view/amplicon-sequencing-using-minion-optimized-from-1d-mhkc34w. Retrieved 27 September 2018.
69. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol 37:540–546. https://doi.org/10.1038/s41587-019-0072-8.
70. Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res 27:737–746. https://doi.org/10.1101/gr.214270.116.
71. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212. https://doi.org/10.1093/bioinformatics/btv351.
72. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421. https://doi.org/10.1186/1471-2105-10-421.
73. Lu J, Salzberg SL. 2018. Removing contaminants from databases of draft genomes. PLoS Comput Biol 14:e1006277. https://doi.org/10.1371/journal.pcbi.1006277.
74. Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Mol Biol Evol 33:1635–1638. https://doi.org/10.1093/molbev/msw046.
75. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods 13:581–583. https://doi.org/10.1038/nmeth.3869.
76. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, Huttley GA, Gregory Caporaso J. 2018. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome 6:90. https://doi.org/10.1186/s40168-018-0470-z.

Downloaded from https://journals.asm.org/journal/mbio on 08 June 2022 by 54.66.212.33.

Chapter 10
Discussion

Fungal infections continue to have high morbidity and mortality rates amongst vulnerable populations despite available antifungal treatments. Especially the increase of rare and uncommon fungal pathogens originating from diverse ecosystems require refined and complex identification tools. As such, rapid and accurate diagnostic systems are essential to reduce morbidity and mortality and improve patient outcomes. Next generation long-read sequencing has a high potential for fungal identification with both metabarcoding and metagenomic approaches. ONT produces a wide variety of long-read sequencers that have read length only limited by the input DNA, allowing large amounts of sequencing data to be generated from one sequencing run. ONT sequencers are scalable for use in large institutions to smaller laboratories with minimal equipment. Implementation of metabarcoding and metagenomic using long-read sequencers in fungal diagnostics requires robust databases of reference barcodes or whole genome sequences, efficient fungal DNA extraction from patient samples, and effective bioinformatic tools suited for fungal identification.

The presented study aimed to contribute to the improvement of metabarcoding and metagenomic long-read sequencing of pathogenic fungal species and hypothesised that doing so would improve fungal diagnostics leading to better patient outcomes.

The first portion of this thesis was focused on improving fungal metabarcoding.

To address the first specific aim: **to improve established databases containing reference fungal barcode sequences for the application of metabarcoding with nanopore sequencing to fungal infections**, improvement of reference databases was done. The ISHAM ITS database was established as a well-curated database of quality-controlled reference ITS sequences for medically relevant fungal species [1]. As a response for the need to improve identification with DNA barcoding, *TEF1α* was proposed as the secondary fungal DNA barcode. However, there was no reliable database to implement the new barcode in

identification [2]. As such, the present study generated 313 primary fungal DNA barcode (ITS) sequences and 297 secondary fungal DNA barcode (*TEF1α*) sequences to expand the repertoire of accurate reference sequences, thus increasing the confidence of identification. The ISHAM ITS database (https://its.mycologylab.org) was then expanded to create the ISHAM Barcoding database to include the secondary fungal DNA barcode, including those sequences generated in this study [3]. Additionally, this study compared the resolution of the primary and secondary barcodes as investigation beyond the establishment of the secondary barcode had not been done. The identification of all analysed taxa was found to be improved with the secondary barcode, supporting the addition of the secondary barcode [4]. The accumulation of this data has improved the databases for reference barcode sequences used in metabarcoding studies.

The second specific aim of the study was **to develop a fungal-specific bioinformatic tool for fungal metabarcoding studies**. Studies identifying fungi to the species level in clinical settings using a metabarcoding approach have been successfully explored [5, 6]. However, the majority of bioinformatic tools have been designed for bacterial 16S metabarcoding and do not consider the complications that arise in the fungal barcodes [7]. Those that are widely available are based upon clustering, such as Pipecraft and QIIME2, or based upon k-mers, such as Kraken2 [8-10]. The present study established a machine learning approach to the classification of fungal species, developed by a convoluted neural network cladogram built upon long amplicons containing the ITS region from nanopore sequencing. The herein established different approach to fungal metabarcoding analysis is beneficial to research, as the wider variety of bioinformatic tools suited to fungal species as tools designed for other microbes will no longer be required.

The remainder of this work then focused on improving fungal nanopore metagenomic studies.

Identification of pathogenic fungal species from patient samples using metagenomic nanopore sequencing was explored in this study to address the third aim: **to investigate the current ability of ONT sequencing technology to identify causative pathogenic fungal species in clinical samples** [5, 6]. Although metabarcoding has successfully been implemented to identify fungal species within clinical samples, the application of metagenomics to such studies is limited. In the presented study, MinION flow cells were used to identify the causative fungal pathogen *P. jirovecii* in all analysed clinical samples. However it was not the most abundant taxa, complicating its application in routine diagnostics [11]. Flongle flow cells are available at a lower price and provide an attractive option for use in routine diagnostics. Flongle flow cells have been successfully applied to fungal metabarcoding studies and metagenomic bacterial identification in a clinical setting [12, 13]. Beside these successful applications for bacterial identification, the present study was unable to identify the causative pathogen in all clinical samples when using the Flongle flow cells. The presented workflow did not allow for optimal use of long-read nanopore sequencing for pathogen identification. As such, the herine carried out studies highlighted the prominent setback of the high abundance of human DNA present in clinical samples and thereby being sequenced. This results in the reduction of identification resolution underlines the need to deplete human DNA in long-read metagenomic studies [14].

The fourth specific aim has focused on **the optimization of DNA extraction and human DNA depletion methods for spiked fungal blood samples to improve sequencing efficiency**, which would overcome the problems accentuated by the results of the previous aim. The requirement to deplete human DNA in sequencing of clinical samples has been previously identified and as such published protocols and commercial kits are available [14-16]. However, these depletion methods have largely been designed for the enrichment of bacterial DNA and their application to samples containing fungi have not been investigated.

This study compared DNA extraction and host depletion methods and presented an optimised DNA extraction and host depletion workflow built upon the modification of a commercial extraction kit and a published depletion protocol [14]. The implementation of these improved the limit of detection of *C. albicans* in spiked blood samples and thus could be beneficial if applied to real clinical samples. As such, his study has built a foundation for further work to reduce the ratio of host to fungal pathogen DNA.

The fifth aim was to **develop a fungal specific bioinformatic pipeline for metagenomic studies**. As outlined throughout this study, tools for long-read nanopore metabarcoding and metagenomic studies have not been developed for the use in fungal identification. The present study determined that bioinformatic tools identified fungal species best when using a quality fungal-specific reference database, emphasising the importance of robust fungal databases [17]. The alignment-based BLAST bioinformatic tool was also found to perform better than the k-mer-based tool and the study also optimized the cut-offs of query coverage of BLAST to improve the performance. This established an optimized workflow using the existing bioinformatic tool BLAST for improved performance in fungal identification from long-read metagenomic and metabarcoding data.

The major limitation of this study was the small sample size. For metagenomic sequencing of clinical samples using MinION and Flongle flow cells, three samples were only tested for each. However, a larger sample size would be ideal. Despite this, the ability of the flow cells to identify pathogens and the need to deplete human cells for efficient and accurate sequencing has clearly been demonstrated. Additionally, in the optimization of DNA extraction and host depletion section, DNA extraction methods were compared by the yield and quality of DNA extracted. However, it would be preferable that extracted DNA would be sequenced to determine which method provided the highest yield of fungal DNA. Also, in this study, there were no repeats for each condition tested due to relocation of the research

laboratory, the results show the general improvement and directions in how optimization can proceed. The present study provides a proof of concept and illuminates the direction of the work still required to apply long-read metabarcoding and metagenomic sequencing to fungal diagnostics.

To progress this field of study, continued work to maintain and update quality controlled fungal barcode databases should be undertaken, particularly targeting species that are underrepresented to provide higher coverage and represent the diversity within each species. Additionally, for metagenomic sequencing, a database of quality fungal whole genome sequences should be established to first contain the prominent fungal pathogens, then being further expanded to include the rarer infection causing species. Further work to optimise DNA extraction and host depletion protocols for different patient sample types should be undertaken to create standardized protocols aimed for use in clinical diagnostics. Potential biases that may arise from host depletion protocols should also be investigated. The bioinformatics tools developed in the current study should be tested against other datasets and further work may be done so they may be used in diagnostics laboratories without requiring specially trained personnel.

Overall summary: The work presented within this thesis has been built upon previous principles and studies in fungal diagnostics. It has progressed the application of long-read metabarcoding and metagenomics to fungal infections by improving reference databases, DNA extraction, and host depletion protocols, and developed fungal specific bioinformatic pipelines. These have formed part of the basis to potentially incorporate long-read sequencing to fungal diagnostics in the future, which will provide a rapid and accurate identification, in combination with antifungal or epidemiological information to result in better patient outcomes and reduce associated hospital costs.

10.1 References

1. Irinyi, L., et al., *International Society of Human and Animal Mycology (ISHAM)-ITS reference DNA barcoding database—the quality controlled standard tool for routine identification of human and animal pathogenic fungi.* Med. Mycol., 2015. **53**(4): p. 313-337.
2. Stielow, J.B., et al., *One fungus, which genes? Development and assessment of universal primers for potential secondary fungal DNA barcodes.* Persoonia, 2015. **35**: p. 242.
3. Meyer, W., et al., *Database establishment for the secondary fungal DNA barcode translational elongation factor 1α (TEF1α).* Genome, 2019. **62**(3): p. 160-169.
4. Hoang, M.T.V., et al., *Dual DNA Barcoding for the Molecular Identification of the Agents of Invasive Fungal Infections.* Front. Microbiol., 2019. **10**: p. 1647.
5. D'Andreano, S., A. Cuscó, and O. Francino, *Rapid and real-time identification of fungi up to species level with long amplicon nanopore sequencing from clinical samples.* Biol. Methods Protoc., 2020. **6**(1): p. 1-6.
6. Ohta, A., et al., *Using nanopore sequencing to identify fungi from clinical samples with high phylogenetic resolution.* Sci. Rep., 2023. **13**(1): p. 9785.
7. Tedersoo, L., et al., *Best practices in metabarcoding of fungi: From experimental design to results.* Mol. Ecol., 2022. **31**(10): p. 2769-2795.
8. Anslan, S., et al., *PipeCraft: Flexible open-source toolkit for bioinformatics analysis of custom high-throughput amplicon sequencing data.* Mol. Ecol. Resour., 2017. **17**(6): p. e234-e240.
9. Bolyen, E., et al., *Reproducible, interactive, scalable and extensible microbiome data science using QIIME2.* Nat. Biotechnol., 2019. **37**(8): p. 852-857.

10. Wood, D.E., J. Lu, and B. Langmead, *Improved metagenomic analysis with Kraken 2.* Genome Biol., 2019. **20**(1): p. 257.

11. Irinyi, L., et al., *Long-read sequencing based clinical metagenomics for the detection and confirmation of Pneumocystis jirovecii directly from clinical specimens: A paradigm shift in mycological diagnostics.* Med. Mycol., 2020. **58**(5): p. 650-660.

12. Tay, D.D., et al., *ITS1 amplicon sequencing of feline gut mycobiome of Malaysian local breeds using Nanopore Flongle.* Arch. Microbiol., 2022. **204**(6): p. 314.

13. Avershina, E., et al., *Ultrafast and Cost-Effective Pathogen Identification and Resistance Gene Detection in a Clinical Setting Using Nanopore Flongle Sequencing.* Front. Microbiol., 2022. **13**: p. 1-10.

14. Charalampous, T., et al., *Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection.* Nat. Biotechnol., 2019. **37**(7): p. 783-792.

15. Marotz, C., et al., *Host DNA Depletion in Saliva Samples for Improved Shotgun Metagenomics.* Methods Mol. Biol., 2021. **2327**: p. 87-92.

16. Marquet, M., et al., *Evaluation of microbiome enrichment and host DNA depletion in human vaginal samples using Oxford Nanopore's adaptive sequencing.* Sci. Rep., 2022. **12**(1): p. 4000.

17. Hu, Y., et al., *Inferring species compositions of complex fungal communities from long- and short-read sequence data.* mBio, 2022. **13**(2): p. 1-16.

Appendix

Appendix 1 Strain List

							Accession Numbers	
Strain ID	**Other Associated ID Numbers**	**Species Name**	**Year Isolated**	**Country of Origin**	**Source**	**Specific Source**	**ITS**	***TEF1α***
WM 03.223	MH4	*Aspergillus flavus*	2003	USA	Clinical		AJ853764.1	
WM 03.225		*Aspergillus flavus*	2003	USA	Clinical		KP068710	
WM 03.230	20156.037	*Aspergillus flavus*	2003	USA			OR195799	
WM 06.64		*Aspergillus flavus*	2006	Australia	Clinical			
WM 06.98		*Aspergillus fumigatus*	2006	Australia	Clinical		EF567971.1	
WM 10.75		*Aspergillus niger*	2010	Oman	Clinical		OR195800	
WM 10.76		*Aspergillus niger*	2010	Oman	Clinical		OR195801	
WM 06.104		*Aspergillus tubingensis*	2004	Australia			KP131636	
WM 07.11	CBS 8244	*Blastobotrys adeninivorans*	1997	Netherlands	Environmental	Soil	EU343811.1	
WM 07.12	CBS 522.75	*Blastobotrys proliferans*	1974	Brazil	Environmental	Nut	FM178324.1	
WM 03.488		*Candida aaseri*	2003	USA			OR195802	
WM 01.157		*Candida albicans*	2001	Australia	Clinical	Blood	EF567987.1	

WM 01.158		*Candida albicans*	2001	Australia	Clinical	Wound	EF567988.1	
WM 02.195		*Candida albicans*	2002	Australia	Clinical	Blood	FM178299.1	
WM 03.584		*Candida albicans*	2003	Australia	Clinical	Blood	KP131672	
WM 229	CBS 562, H6.1, MH 99	*Candida albicans*	1923	Uruguay	Clinical	Skin of man with interdigital mycosis	EF567995.1	
WM 230	CBS 1905, MH 101	*Candida albicans*	1923	Netherlands	Clinical		AJ853768.1	
WM 231	CBS 5983, H7.1.1, MH 100	*Candida albicans*			Clinical		EF568101.1	
WM 07.90	CBS 1898 (T), MUCL 29808	*Candida blankii*		Canada	Veterinary	Blood of *Putorius vison* (mink)	EU343873.1	
WM 07.20	CBS 6420 (T), MUCL 29809	*Candida boleticola*	1971		Environmental	Fruiting body of *Astraeus hygrometricus* (toadstool)	OR195803	
WM 07.23	CBS 5719 (T), MUCL 29815	*Candida chilensis*	1961	Chile	Environmental	Rotten stump of *Nothofagus* sp.	FM178354.1	
WM 07.91	CBS 8178 (T), MUCL 29818	*Candida coipomensis*	1980	Chile	Environmental	Rotten log of *Eucryphia cordifolia*	FM178314.1	

WM 07.25	MUCL 44064	*Candida diddensiae*	2007	Belgium			FM178307.1	
WM 07.24	CBS 2214 (T), MUCL 29822	*Candida diddensiae*		USA	Environmental	Shrimp		
WM 02.73		*Candida dubliniensis*	2002	Australia	Clinical	Peripheral	EF567996.1	
WM 03.119		*Candida dubliniensis*	2002	Australia	Clinical	Central venous line	EF567997.1	
WM 03.133		*Candida dubliniensis*	2002	Australia	Clinical	Human blood	FM178301.1	
WM 03.160		*Candida dubliniensis*	2001	Australia	Clinical	Peripheral	EF567998.1	
WM 03.78		*Candida dubliniensis*	2002	Australia	Clinical	Peripheral	FM178300.1	
WM 03.79		*Candida dubliniensis*	2002	Australia	Clinical	Peripheral	EF568000.1	
WM 07.28	CBS 6160 (T), MUCL 30033	*Candida entomophila*		South Africa	Environmental	Tunnel of *Crossotarsus externedentatus*, in *Ficus sycomorus*	FM178368.1	
WM 07.33	CBS 6380 (T), MUCL 29830	*Candida fructus*	1962	Japan	Environmental	Fruit of *Musa sapientum* (banana)	EU343830.1	
WM 07.34	CBS 6381 (T), MUCL 29862	*Candida fructus*	1997	Japan	Environmental	Fruit of *Musa sapientum* (banana)	EU343831.1	

WM 02.57		*Candida glabrata*	2002	Australia	Clinical	Groin artery	EF568002.1	
WM 03.500		*Candida glabrata*	2003	USA			KP131711	
WM 07.36	CBS 5691 (T), MUCL 29834	*Candida glaebosa*		Japan	Veterinary	Fish	EU343832.1	
WM 899	CIN 01	*Candida haemulonis*		New Zealand			OR195804	
WM 902	CIN 04	*Candida haemulonis*		New Zealand			OR195805	
WM 07.40	CBS 5604 (T), MUCL 29843	*Candida incommunis*		Japan	Environmental	Grape must	EU343836.1	
WM 07.41	MUCL 43783	*Candida incommunis*		Belgium			EU343837.1	
WM 11	CBS 180 (T), MH 18	*Candida inconspicua*		Netherlands	Clinical	Human sputum	AJ853766.1	
WM 873		*Candida inconspicua*		New Zealand				
WM 900		*Candida inconspicua*		New Zealand				
WM 901	CIN 03	*Candida inconspicua*		New Zealand			OR195806	
WM 904	UN 01	*Candida inconspicua*		New Zealand			OR195807	
WM 905	UN 02	*Candida inconspicua*		New Zealand			OR195808	
WM 906		*Candida inconspicua*		New Zealand				

WM 07.42	CBS 6036 (T), MUCL 29970	*Candida insectalens*	1969	South Africa	Environmental	Tunnel of *Platypus externedentatus*, in *Ficus sycomorus*	EU364562.1	
WM 07.44	CBS 6213 (T), MUCL 29915	*Candida insectorum*		South Africa	Environmental	Frass of *Phorocantha recurva*, in *Eucalyptus maculata*	OR196116	
WM 03.499		*Candida intermedia*		USA			KP131723	
WM 811	CBS 572 (T), MH 20	*Candida intermedia*		Puerto Rico	Clinical	Faeces of American woman with tropical sprue	EF568011.1	
WM 07.50	CBS 1952 (T), MUCL 29859	*Candida membranifaciens*			Clinical	Urine	EU343843.1	
WM 07.53	CBS 2095, MUCL 29949	*Candida mesenterica*		Germany	Environmental	Beer	EU343846.1	
WM 01.56	PL 448, CBS 10907 (T), ATCC 96144	*Candida metapsilosis*		USA	Clinical	Human hand	KP131735	
WM 04.400		*Candida metapsilosis*	2004	Australia	Clinical	Catheter	KP131737	
WM 815	CBS 7409 (T)	*Candida mucifera*	1997	Brazil	Veterinary	Toad liver	EF568082.1	
WM 07.55	CBS 6524 (T), MUCL 30042	*Candida multigemmis*		United Kingdom	Environmental	Fruit of *Rubus idaeus* (raspberry)	FR819711.1	

WM 19	CBS 4239 (T), MH 19	*Candida norvegica*		Norway	Clinical	Human sputum		
WM 03.414		*Candida orthopsilosis*	2003	Australia	Clinical	Central venous line	KP131748	
WM 04.53		*Candida orthopsilosis*	2002	Australia	Clinical	Human blood	KP068749	
WM 04.54		*Candida orthopsilosis*	2005	Australia	Clinical		KP068750	
WM 02.130a		*Candida parapsilosis*	2001	Australia	Clinical	Port-a-cath	EF568027.1	
WM 02.166		*Candida parapsilosis*	2001	Australia	Clinical	Human blood	EF568028.1	
WM 02.177		*Candida parapsilosis*		Australia	Clinical		EF568029.1	
WM 02.200		*Candida parapsilosis*	2002	Australia	Clinical		EF568030.1	
WM 02.61b		*Candida parapsilosis*	2002	Australia	Clinical	Peripheral	EF568032.1	
WM 02.75		*Candida parapsilosis*	2002	Australia	Clinical	Human Sputum	EF568033.1	
WM 02.89		*Candida parapsilosis*	2001	Australia	Clinical	Peripheral	EF568034.1	
WM 02.95		*Candida parapsilosis*	2002	Australia	Clinical	Yellow lumen Hickmans	EF568035.1	
WM 09.77	2009:4:7A	*Candida parapsilosis*	2009				KP131800	
WM 17.12		*Candida parapsilosis*	2017	Australia	Clinical	Eye cornea	OR196117	
WM 17.13		*Candida parapsilosis*	2017	Australia	Clinical	Eye cornea	OR196118	

WM 17.14		*Candida parapsilosis*	2017	Australia	Clinical	Eye cornea	OR196119	
WM 17.15		*Candida parapsilosis*	2017	Australia	Clinical	Eye cornea	OR196120	
WM 17.16		*Candida parapsilosis*	2017	Australia	Clinical	Eye cornea	OR196121	
WM 67		*Candida parapsilosis*	1996				FM178294.1	
WM 07.1	CBS 8164 (T)	*Candida railenensis*	1980	Chile	Environmental	Rotten trunk of *Nothofagus dombeyii*	EU343803.1	
WM 07.59	CBS 5838 (T), MUCL 29936	*Candida santamariae*	1959		Environmental	*Hymenochaete rubiginosa*	EU343850.1	
WM 07.60	MUCL 29806	*Candida santamariae*	2007	Belgium	Environmental	Cider	FM178312.1	
WM 07.61	CBS 6563 (T), MUCL 29883	*Candida savonica*	1972	Finland	Environmental	*Betula* sp. (birch tree)	EU343852.1	
WM 07.62	MUCL 31357	*Candida savonica*	2007	Belgium			EU343853.1	
WM 07.63	CBS 8118 (T), MUCL 29886	*Candida sequanensis*		France	Environmental	Grain of *Avena* sp. (oats)	EU343854.1	
WM 07.64	CBS 6277 (T), MUCL 29938	*Candida silvatica*		South Africa	Environmental	Tunnel of *Crossotarsus externedentatus*, in *Ficus sycomorus*		

WM 07.65	CBS 8175 (T), MUCL 29889	*Candida sophiae-reginae*	1980	Chile	Environmental	Rotten trunk of *Laurelia sempervirens*	EU343856.2	
WM 07.66	MUCL 47126	*Candida sophiae-reginae*	2007	Belgium	Environmental	Brewery	FM178355.1	
WM 07.69	CBS 7422 (T), MUCL 34510	*Candida tanzawaensis*	1966	Japan	Environmental	*Polytrichum commune* (moss)	EU343858.1	
WM 223	DUMC M0069	*Candida tropicalis*	1993	Australia				
WM 01.203		*Candida tropicalis*	2001	Australia	Clinical	Urine	EF568038.1	
WM 02.181		*Candida tropicalis*	2001	Australia	Clinical	Peripheral	EF568039.1	
WM 02.86		*Candida tropicalis*	2002	Australia	Clinical	Peripheral	EF568040.1	
WM 10.103		*Candida tropicalis*	2010	Oman	Clinical	Urine		
WM 10.107		*Candida tropicalis*	2010	Oman	Clinical	Human sputum		
WM 233	CBS 94 (T), K19, MH45, ATCC 750	*Candida tropicalis*		Australia	Clinical	Bronchitic patient	EF568042.2	
WM 03.458		*Candida zeylanoides*	2003	USA			KP131820	
WM 20		*Candida zeylanoides*		Norway	Clinical	Human sputum	DQ249203.1	
WM 64	CBS 619 NT, MH 51	*Candida zeylanoides*			Clinical	Case of blastomycotic macrogloss	DQ249202.1	

WM 284	CBS 864	*Citeromyces matritensis*		Netherlands	Environmental	Sweetened condensed milk	KP131823	
WM 10.51		*Cladosporium sphaerospermum*	2010	Oman	Clinical		OR196122	
WM 02.466	CBS 5094	*Clavispora lusitaniae*		Germany	Clinical	Urine	EF568047.1	
WM 02.479	CBS 5901 (T)	*Clavispora lusitaniae*	1979		Environmental	Droppings of *Corvus albus* (crow)	EF568048.1	
WM 02.92		*Clavispora lusitaniae*	2001	Australia	Clinical	Peripheral	DQ249208.2	
WM 18	CBS 4413 (T), K64.1	*Clavispora lusitaniae*		Portugal	Environmental	Caecum of pig	EF568024.1	
WM 35	CBS 6936 (T), K65.1, MH 52	*Clavispora lusitaniae*		Israel	Environmental	Citrus essence	EF568049.1	
WM 01.1	NIH 37-40-40, 46914	*Cryptococcus gattii*	1994	Thailand	Clinical	Cerebrospinal fluid	KP132170	
WM 01.128	ATCC 32609, NIH 444, JH 818, CBS 6956	*Cryptococcus gattii*	1975	USA	Clinical	Human sputum	AB087656.1	
WM 01.2	NIH 37-362-28, IFM46916	*Cryptococcus gattii*	1994	Thailand	Clinical	Cerebrospinal fluid	AB087642.1	
WM 01.3	NIH 38-1539-59, IFM46919	*Cryptococcus gattii*	1995	Thailand	Clinical	Cerebrospinal fluid	AB087645.1	

WM 02.138	NIH 112, ATCC 24065	*Cryptococcus gattii*		USA	Clinical		AB087622.1	
WM 02.209	F 3179	*Cryptococcus gattii*	2001	Canada	Clinical	Sputum	OR196123	
WM 02.35	CDC R268	*Cryptococcus gattii*	2001	Canada	Clinical	Cerebrospinal fluid	OR196124	
WM 02.48	CDC R406	*Cryptococcus gattii*		Canada	Clinical	Cerebrospinal fluid	OR196125	
WM 03.27		*Cryptococcus gattii*	1999	Australia	Environmental	*Eucalyptus camaldulensis*	AJ493568.1	
WM 03.697		*Cryptococcus gattii*	2003	Canada	Veterinary		KP068807	
WM 05.414		*Cryptococcus gattii*	2005	Brazil	Veterinary		JN944027.1	
WM 05.415		*Cryptococcus gattii*	2005	Brazil	Clinical		JN944026.1	
WM 05.423	IFM 5838	*Cryptococcus gattii*		USA			AB034645.1	
WM 05.424	IFM 5883, UFPR 2527	*Cryptococcus gattii*		Brazil	Clinical		AB087631.1	
WM 05.425	IFM 46566, C61	*Cryptococcus gattii*		Costa Rica	Clinical		AB087638.1	
WM 05.427	IFM 46915, 37-140-140	*Cryptococcus gattii*		Thailand	Clinical		AB087641.1	
WM 05.429	IFM 46917, 37-381-47	*Cryptococcus gattii*		Thailand	Clinical		AB087643.1	
WM 05.430	IFM 46918, 37-1852-23	*Cryptococcus gattii*		Thailand	Clinical		AB087644.1	
WM 05.432	IFM 47258, F-1623	*Cryptococcus gattii*		Japan	Clinical		AB087646.1	

WM 05.433	IFM 48213, F T84	*Cryptococcus gattii*		Brazil	Clinical		AB087647.1	
WM 05.434	IFM 48214, FT 78	*Cryptococcus gattii*		Brazil			AB087648.1	
WM 05.435	IFM 48216, FT 119	*Cryptococcus gattii*		Brazil	Clinical		AB087649.1	
WM 05.436	IFM 48218, BV 33	*Cryptococcus gattii*		Venezuela	Clinical		AB087650.1	
WM 05.437	IFM 48221, A288	*Cryptococcus gattii*		Brazil	Clinical		AB087652.1	
WM 05.438	IFM 48222, FT 102	*Cryptococcus gattii*		Brazil	Clinical		AB087653.1	
WM 05.439	IFM 48297, FT 130	*Cryptococcus gattii*		Brazil	Clinical		AB087654.1	
WM 05.440	IFM 48302, BV 5	*Cryptococcus gattii*		Venezuela	Clinical		AB087655.1	
WM 05.77	AV55, CBS 10090	*Cryptococcus gattii*	1998	Greece	Clinical			
WM 06.13	CBS 7750, 451-SF13.2	*Cryptococcus gattii*	1990	USA	Environmental	*Eucalyptus camaldulensis*	AB087658.1	
WM 06.4	ATCC 32609, NIH 444, CBS 6955	*Cryptococcus gattii*	1975	USA	Clinical	Sputum	OR196126	
WM 06.604	IFM 45714, S-D2Y, LY-7	*Cryptococcus gattii*		China	Clinical		AB087637.1	
WM 06.605	IFM 45713	*Cryptococcus gattii*		China	Clinical		AB087636.1	
WM 06.635	DMST 20766	*Cryptococcus gattii*	1994	Thailand	Clinical	Cerebrospinal fluid	OR196127	

WM 06.636	DMST 20767	*Cryptococcus gattii*	1995	Thailand	Clinical		OR196128	
WM 05.459		*Cryptococcus gattii*		Brazil	Clinical		JN944024.1	
WM 11.115		*Cryptococcus gattii*	2007	Colombia	Clinical	Cerebrospinal fluid	KP068851	
WM 11.128	JS99	*Cryptococcus gattii*		USA				
WM 11.138	JS109	*Cryptococcus gattii*		USA			OR196129	
WM 11.36	JS85	*Cryptococcus gattii*		USA			OR196130	
WM 11.38	JS87	*Cryptococcus gattii*		USA			OR196131	
WM 11.41	JS92	*Cryptococcus gattii*		USA			OR196132	
WM 11.938	B8262, MAS92-409	*Cryptococcus gattii*		USA	Clincical		KP068858	
WM 1220	CBS 7748	*Cryptococcus gattii*	1989	Australia	Environmental	*Eucalyptus camaldulensis*		
WM 1219	CBS 7747	*Cryptococcus gattii*	1989	Australia	Environmental	*Eucalyptus camaldulensis*	AB087672.1	
WM 1221	CBS 7749	*Cryptococcus gattii*	1989	Australia	Environmental	*Eucalyptus camaldulensis*	AB034647.1	
WM 175	CBS 10081, 91B3214, Harkness,	*Cryptococcus gattii*	1992	USA	Environmental	*Eucalyptus* spp.	KP068818	

	689 | F 07, WM 161							
WM 178	49435, Colter, CBS 10082, IFM 50894	*Cryptococcus gattii*	1991	Australia	Clinical	Human lung	AJ493567.1	
WM 455	WGH-1, R501883, HT | IFM50897	*Cryptococcus gattii*	1994	Australia	Clinical	Cerebrospinal fluid	AB087668.1	
WM 477	HEC 11102, LMM 21	*Cryptococcus gattii*	1994	Brazil	Clinical	Urine	AJ493569.1	
WM 830		*Cryptococcus gattii*	1992	Papua New Guinea	Clinical		AJ493563.1	
WM 834		*Cryptococcus gattii*		Papua New Guinea	Clinical		KP068803	
WM 08.311		*Cryptococcus gattii*	1996	Australia	Veterinary	Dog (Dalmatian), cranial mediastinal mass	KP068845	
WM 01.4	CUH-16, IFM 5817	*Cryptococcus neoformans var. grubii*		Japan	Clinical		AB087812.1	
WM 05.406	LMM 192	*Cryptococcus neoformans var. grubii*		Brazil	Clinical	Sputum	OR196133	
WM 05.490		*Cryptococcus neoformans var. grubii*	2000	Peru	Clinical	Cerebrospinal fluid		

WM 05.523		*Cryptococcus neoformans var. grubii*	2001	Peru	Clinical	Cerebrospinal fluid		
WM 05.464	LMM 616	*Cryptococcus neoformans var. grubii*		Brazil	Clinical			
WM 05.473		*Cryptococcus neoformans var. grubii*	1997	Peru	Clinical	Cerebrospinal fluid	KP068864	
WM 05.483		*Cryptococcus neoformans var. grubii*	2000	Peru	Clinical	Cerebrospinal fluid	KP068893	
WM 12.17		*Cryptococcus neoformans var. grubii*	2001	Botswana	Clinical	Cerebrospinal fluid	KP068901	
WM 12.127		*Cryptococcus neoformans var. grubii*	2012	Italy	Veterinary	Cat nasal swab	KP068887	
WM 01.166		*Cryptococcus neoformans var. grubii*	2000	Brazil	Clinical	Cerebrospinal fluid	OR196134	
WM 06.689		*Cryptococcus neoformans var. grubii*	2004	Thailand	Clinical	Cerebrospinal fluid	KP068899	
WM 10.172		*Cryptococcus neoformans var. grubii*	2010	USA			KP068884	
WM 12.23		*Cryptococcus neoformans var. grubii*		Botswana	Clinical	Cerebrospinal fluid	OR196135	

WM 628		*Cryptococcus neoformans var. grubii*	1988	Australia	Clinical	Cerebrospinal fluid	OR196136	
WM 13.46		*Cryptococcus neoformans var. grubii*		Italy	Clinical	Cerebrospinal fluid	KP068891	
WM 148		*Cryptococcus neoformans var. grubii*	1989	Australia	Clinical	Cerebrospinal fluid	AJ493551.1	
WM 856		*Cryptococcus neoformans var. grubii*		Australia	Clinical		KP132182	
WM 09.110		*Cryptococcus neoformans var. grubii*					OR197580	
WM 01.126		*Cryptococcus neoformans var. neoformans*		USA			OR196137	
WM 04.171		*Cryptococcus neoformans var. neoformans*	1999	Thailand	Environmental	Pigeon Droppings	OR196138	
WM 04.172		*Cryptococcus neoformans var. neoformans*		Germany	Clinical		AJ493560.1	
WM 12.178		*Cryptococcus neoformans var. neoformans*		Italy	Veterinary	Cat nasal swab	KP068907	
WM 13.4		*Cryptococcus neoformans var. neoformans*		Italy	Clinical	Cerebrospinal fluid		
WM 56	CBS 570 (T)	*Cutaneotrichosporon curvatum*	1932	Netherlands	Clinical	Human sputum	KP131901	

WM 45		*Cyberlindnera jadinii*	1996	Australia			DQ249199.2	
WM 07.14		*Danielozyma litseae*	2001	India	Environmental	Insect frass on *Litsea polyantha*	EU343814.1	
WM 07.13		*Danielozyma ontarioensis*	1998	Canada	Enviroenmtnal	Frass from tunnels of insect larvae, in dead *Picea* sp. (spruce tree)	FM178337.1	
WM 6		*Diutina catenulata*		Puerto Rico	Clinical	Faeces of man with dysentry	KP131694	
WM 04.488		*Eurotium amstelodami*	2004	Australia	Clinical		AJ853761.1	
WM 13.103		*Filobasidium magnum*	2012	Italy	Veterinary	Cat nasal swab	KP131912	
WM 13.104		*Filobasidium magnum*	2012	Italy	Veterinary	Cat nasal swab	KP131913	
WM 13.86		*Filobasidium magnum*	2012	Italy	Veterinary	Cat nasal swab	KP068785	
WM 13.87		*Filobasidium magnum*	2012	Italy	Veterinary	Cat nasal swab	KP068786	
WM 13.98		*Filobasidium magnum*	2012	Italy	Veterinary	Cat nasal swab	KP131914	
WM 07.185		*Fusarium proliferatum*	2007	Australia	Clinical		KP068933	
WM 07.197		*Fusarium solani*	2007	Australia	Clinical	Nails	KP068934	

WM 17.22		*Geotrichum candidum*		Australia			OR196139	
WM 17.20		*Geotrichum candidum*		Australia			OR196140	
WM 17.23		*Geotrichum candidum*		Australia			OR196141	
WM 17.24		*Geotrichum candidum*		Australia			OR196142	
WM 05.217	22.12	*Geotrichum candidum*	2005	Australia			OR196144	
WM 07.304		*Geotrichum candidum*	2007	Australia	Clinical		KP132256	
WM 07.58		*Groenewaldozyma salmanticensis*		Spain	Environmental	Alpichin	EU343849.1	
WM 775		*Hannaella luteola*	1997	Japan	Environmental	Atmosphere	KP132260	
WM 794		*Hyphopichia burtonii*	1998		Environmental	Pollen, carried by wild bees		
WM 07.39		*Hyphopichia homilentoma*		South Africa	Environmental	Tunnel of *Sinoxylon*, in *Combretum apiculatus*	EU343835.1	
WM 07.56		*Hyphopichia rhagii*	1960	Germany	Veterinary	*Harpium inquisitor* (cerambycid beetle), female	EU343848.1	
WM 03.556(1)		*Inopinatum lactosum*		USA	Clinical		OR196153	
WM 28		*Kluyveromyces lactis*	1996	Australia			OR196143	

WM 03.267		*Kluyveromyces marxianus*		USA			KP132329	
WM 03.289		*Kluyveromyces marxianus*		USA			KP132330	
WM 03.295		*Kluyveromyces marxianus*		USA			KP132331	
WM 03.297		*Kluyveromyces marxianus*		USA			KP132332	
WM 04.198		*Kluyveromyces marxianus*	2004	Australia	Clinical		KP132333	
WM 04.263		*Kluyveromyces marxianus*	2004	Australia	Clinical		KP132334	
WM 04.305		*Kluyveromyces marxianus*	2004	Australia	Clinical		KP132335	
WM 13		*Kluyveromyces marxianus*		Netherlands	Environmental	Kefyr grain	KP132338	
WM 39		*Kluyveromyces marxianus*	1998				EF568057.1	
WM 01.20		*Kodamaea ohmeri*	2001	Chile	Environmental	Bird Guano	FM178297.1	
WM 10.200		*Kodamaea ohmeri*	2010	Australia	Environmental	Small hive beetle (*Aethina tumida*) newly emerged adults	KP132359	
WM 10.210		*Kodamaea ohmeri*	2010	Australia	Environmental	Adult small hive beetle (*Aethina*		

						tumida), mixed sex		
WM 10.220		*Kodamaea ohmeri*	2010	Australia	Environmental	Adult small hive beetle (*Aethina tumida*), mixed sex	KP132360	
WM 806	CBS 1950 (T), H27.1.1, MH 15	*Kodamaea ohmeri*	1997		Environmental	Either figs or dates	FM178323.1	
WM 807	CBS 5367 (T)	*Kodamaea ohmeri*	1998	USA	Environmental	Cucumber brine	EF568058.1	
WM 796	CBS 1993 (T), MH 73	*Kuraishia capsulata*	1994	Canada	Environmental	Insect frass on conifer	EF568066.1	
WM 07.2	CBS 7702 (T)	*Kurtzmaniella fragi*	1962	Japan	Environmental	Rotting fruit of *Fragaria chiloensi var. ananassa* (strawberry)	EU343804.1	
WM 17.17		*Magnusiomyces capitatus*		Australia			OR196145	
WM 17.19		*Magnusiomyces capitatus*		Australia			OR196146	
WM 17.25		*Magnusiomyces capitatus*		Australia			OR196147	
WM 17.26		*Magnusiomyces capitatus*		Australia			OR196148	
WM 07.38	CBS 9146 (T), MUCL 46200	*Metschnikowia hawaiiana*		USA	Environmental	Flower of *Ipomoea indica*	EU343834.1	

WM 799		*Metschnikowia reukaufii*					OR196149	
WM 02.131		*Meyerozyma guilliermondii*	2002	Australia	Clinical	Peripheral	EF568005.1	
WM 02.132		*Meyerozyma guilliermondii*	2002	Australia	Clinical	Central venous line	FM178298.1	
WM 02.361		*Meyerozyma guilliermondii*	2001	Australia	Clinical	Peripheral	EF568006.1	
WM 02.72		*Meyerozyma guilliermondii*	2002	Australia	Clinical	Peripheral	EF568008.1	
WM 02.91		*Meyerozyma guilliermondii*	2001	Australia	Clinical	Peripheral	DQ249193.1	
WM 827	CBS 2030 (T), K78.1, MH 76	*Meyerozyma guilliermondii*	1997	USA	Environmental	Insect frass on *Ulmas americana*	EF568003.1	
WM 802	CBS 2020 (T), MH 7	*Millerozyma farinosa*	1997	Trinidad and Tobago	Environmental	Fermenting cacao	AJ853775.1	
WM 803	CBS 185 (T), MH 74	*Millerozyma farinosa*	1889	Poland	Environmental	Jopen beer	EF568067.1	
WM 13.126		*Naganishia albida*			Veterinary	Italy	KP131887	
WM 13.128		*Naganishia albida*	2012	Italy	Veterinary	Cat nasal swab	KP068780	
WM 13.134		*Naganishia albida*			Veterinary	Italy	KP131888	
WM 13.122		*Naganishia albida*			Veterinary	Italy	KP131885	
WM 13.125		*Naganishia albida*			Veterinary	Italy	KP131886	

WM 282		*Naganishia albida*		Italy	Clinical	Erythamatous desquamative skin lesion	KP131890	
WM 07.30	CBS 1737, MUCL 29961	*Nakazawaea ernobii*	2007	Belgium			OR196150	
WM 07.31	CBS 5214 (T), MUCL 30037	*Nakazawaea ernobii*	1998		Environmental	Mycetome of larva of *Ernobius abietis*	EU343828.1	
WM 797	CBS 4140 (T), MH 77	*Nakazawaea holstii*	1950	Canada	Environmental	Frass on *Picea* sp. (yellow spruce)	EF568068.1	
WM 13.136		*Papiliotrema aurea*	2012	Italy	Veterinary	Cloacal swab, bird	KP131893	
WM 02.495		*Papiliotrema laurentii*	2002	Australia	Clinical		KP131906	
WM 857		*Papiliotrema laurentii*	2000	Australia	Clinical		KP131910	
WM 06.341		*Penicillium chrysogenum*	2004	Australia	Environmental		DQ249212.1	
WM 06.338	AS-80	*Penicillium brevicompactum*		Australia	Clinical		OR196151	
WM 02.78		*Pichia kudriavzevii*	2002	Australia	Clinical	Central urine	EF568012.1	
WM 03.103		*Pichia kudriavzevii*	2002	Australia	Clinical	PICC line	EF568013.1	
WM 03.189		*Pichia kudriavzevii*	2003	Australia	Clinical	Peripheral	KP132512	

WM 03.190		*Pichia kudriavzevii*	2003	Australia	Clinical	Peripheral	EF568015.1	
WM 03.204		*Pichia kudriavzevii*	2003	Australia	Clinical	Human left arm	EF568016.1	
WM 03.98		*Pichia kudriavzevii*	2002	Australia	Clinical	Central venous line	EF568017.1	
WM 14	CBS 573 (T), H24.1.1, MH 22	*Pichia kudriavzevii*	1996	Sri Lanka	Clinical	Human sputum, of bronchitic convict	AF246989.1	
WM 32	CBS 638 (T), H3.1	*Pichia membranifaciens*	1997	Germany	Enrionmental	Wine	FM178293.1	
WM 46	CBS 107 (T)	*Pichia membranifaciens*					OR196152	
WM 872	CV 01, MH 25	*Pichia membranifaciens*		New Zealand				
WM 05.23		*Rhodotorula mucilaginosa*	2005	Australia	Clinical		KP132587	
WM 09.204		*Rhodotorula mucilaginosa*	2009	Australia	Clinical		KP068980	
WM 11.927		*Rhodotorula mucilaginosa*	2011	Australia	Clinical		KP132588	
WM 07.26	CBS 4074 (T), MUCL 29905	*Saturnispora diversa*		Japan	Environmental	Grape must	FM178308.1	
WM 07.27	MUCL 40655	*Saturnispora diversa*	1997	Thailand	Environmental	Flower	FR819717.1	
WM 06.504		*Scedosporium apiospermum*	2004	Australia	Clinical	Human ear		
WM 08.270	FMR 8983 | 138016	*Scedosporium aurantiacum*	2008	Spain	Clinical		OR196154	

WM 09.108	P09/1059598 Y, 1847	*Scedosporium aurantiacum*	2009	Australia	Clinical		OR196155	
WM 04.497		*Scedosporium aurantiacum*		Australia	Clinical	Sphenoid sinus tissue	AJ853740.1	
WM 06.385		*Scedosporium aurantiacum*	2001	Australia	Clinical	Human ear swab	KP068990	
WM 06.480		*Scedosporium aurantiacum*	2004	Australia	Clinical	Human sputum, bronchial wash		
WM 06.528		*Scedosporium aurantiacum*	2004	Australia	Clinical	Human sputum, bronchial wash		
WM 06.467		*Scedosporium aurantiacum*	2004	Australia	Clinical	Human bone tissue	KP132687	
WM 06.477		*Scedosporium boydii*	2004	Australia	Clinical	Human ear	KP068984	
WM 09.122	M2, IHEM 15933, CBS 101.22	*Scedosporium boydii*	2009				OR196156	
WM 07.29	CBS 6248 (T), MUCL 30034	*Scheffersomyces ergatensis*		Spain	Environmental	Larva of *Ergates faber*, in stump of *Pinus* sp.	EU343826.1	
WM 09.205		*Starmerella apicola*	2009	Australia	Environmental		KP131673	
WM 07.6	CBS 8846 (T)	*Suhomyces canberraensis*					FM178344.1	

WM 07.8	CBS 8847 (T)	*Suhomyces prunicola*					EU343810.1	
WM 07.7	CBS 8848 (T)	*Suhomyces tanzawaensis*	2001				EU343809.1	
WM 820	CBS 133 T	*Torulaspora delbrueckii*		Netherlands			OR196157	
WM 821	CBS 1146 NT	*Torulaspora delbrueckii*	1904				EF568083.1	
WM 1069		*Trichomonascus ciferrii*	2000	Australia	Clinical		KP132795	
WM 908A		*Trichomonascus ciferrii*	2000	Australia	Clinical		KP132796	
EM 06.879		*Trichophyton rubrum*	2006	Australia	Clinical		KP069000	
WM 06.351		*Trichophyton rubrum*	2003	Australia	Clinical	Nails		
WM 04.474		*Trichophyton rubrum*	2004	Australia	Clinical	Human skin, foot		
WM 03.423		*Trichosporon asahii*	2003	Australia	Clinical	Central venous line tip	KP069015	
WM 1000		*Trichosporon asahii*	2000	Australia	Clinical	Oral swab	KP069016	
WM 03.421	1-2010-0556 A, 56x	*Trichosporon asahii*	2003	Australia	Clinical		OR196158	
WM 269	CBS 2041, H2.1.1, H4.1, MH 54	*Vanrija humicola*	1981				EF568050.1	
WM 13.106		*Vishniacozyma carnescens*	2012	Italy	Veterinary	Cloacal swab, bird	KP131895	

WM 13.114		*Vishniacozyma carnescens*	2012	Italy	Veterinary	Cloacal swab, bird	KP131898	
WM 13.115		*Vishniacozyma carnescens*	2012	Italy	Veterinary	Cloacal swab, bird	KP131899	
WM 13.116		*Vishniacozyma carnescens*	2012	Italy	Veterinary	Dropping swab, bird	KP131900	
WM 13.107	2.82, 205	*Vishniacozyma carnescens*		Italy	Veterinary		OR196159	
WM 13.108		*Vishniacozyma carnescens*	2012	Italy	Veterinary	Cloacal swab, bird	KP131896	
WM 13.110		*Vishniacozyma carnescens*	2012	Italy	Veterinary	Cloacal swab, bird	KP131897	
WM 13.111		*Vishniacozyma carnescens*	2012	Italy	Veterinary	Cat nasal swab	KP068783	
WM 02.251		*Wickerhamomyces anomalus*	2002	Australia	Clinical	Human blood	KP132888	
WM 03.505		*Wickerhamomyces anomalus*	2003	USA			KP132891	
WM 03.506		*Wickerhamomyces anomalus*	2003	USA			KP132892	
WM 03.507		*Wickerhamomyces anomalus*	2003	USA			KP132893	
WM 03.508		*Wickerhamomyces anomalus*	2003	USA			KP132894	
WM 03.509		*Wickerhamomyces anomalus*	2003	USA			KP132895	
WM 03.511		*Wickerhamomyces anomalus*	2003	USA			KP132896	
WM 03.512		*Wickerhamomyces anomalus*	2003	USA			KP132897	

WM 03.72		*Wickerhamomyces anomalus*	2002	Australia	Clinical	Peripheral	KP132904	
WM 07.51	MUCL 30392	*Wickerhamomyces anomalus*		Belgium			OR196160	
WM 824	CBS 5759 NT, MH 71	*Wickerhamomyces anomalus*	1997				DQ249196.1	
WM 03.501		*Wickerhamomyces anomalus*	2003	USA			KP132890	
WM 2194		*Wickerhamomyces anomalus*		Australia			FM178296.1	
WM 825	CBS 605 (T)	*Wickerhamomyces anomalus*	2008	Italy	Clinical	Pus from lung from dead tuberculosis patient	DQ249195.2	
WM 07.45	MUCL 37920	*Yamadazyma Mexicana*	2007	Belgium			EU343839.1	
WM 804	CBS 5815, MH 48	*Yamadazyma Mexicana*	1997	Italy	Environmental	Grape must	FM178295.1	
WM 805	CBS 7066 (T), MH 80	*Yamadazyma Mexicana*	1997	Mexico	Environmental	Soft rot of *Stenocereus gummosis*	EF568069.1	
WM 06.835		*Yamadazyma scolyti*	2003	Canada	Environmental	Lodgepole pines infected with mountain pine	FR819713.1	
WM 06.386		*Yamadazyma scolyti*	2003	Canada	Environmental	Lodgepole pines infected with mountain pine	FR819712.1	

WM 07.5	CBS 4802 (T)	*Yamadazyma scolyti*	1997	USA	Environmental	Frass of *Scolytus ventralis* (bark beetle) in *Abies concolor* (red fir)	EU343807.1	
WM 07.4	CBS 9722 (T)	*Yarrowia galli*	2004	USA	Veterinary	Chicken liver	EU343806.1	
WM 01.192	MMRL 222, H31.1, MH 23	*Yarrowia lipolytica*	1993	USA	Clinical	Human right leg	EF568020.1	
WM 01.197	MMRL 225	*Yarrowia lipolytica*	1993	USA	Clinical	Human tongue	KP132917	
WM 01.198	MMRL 233	*Yarrowia lipolytica*	1993	USA	Clinical	Human blood	KP132918	
WM 01.204	MMRL 238	*Yarrowia lipolytica*	1993	USA	Clinical		KP132920	
WM 01.297	MMRL 248	*Yarrowia lipolytica*	1993	USA	Clinical	Human blood	KP132921	
WM 01.211	MMRL 262	*Yarrowia lipolytica*	1993	USA	Clinical	Stool	KP132922	
WM 03.235		*Yarrowia lipolytica*	2003	Australia	Clinical	Central venous line	KP132923	
WM 03.240		*Yarrowia lipolytica*		USA			KP132924	
WM 03.249		*Yarrowia lipolytica*		USA			KP132925	
WM 03.254		*Yarrowia lipolytica*		USA			KP132926	

WM 03.255		*Yarrowia lipolytica*		USA			KP132927	
WM 1041		*Yarrowia lipolytica*	2000	Australia	Clinical		KP132929	
WM 17	CBS 6124 (T)	*Yarrowia lipolytica*	1945	USA	Environmental	Maize Processing Plant	DQ249205.2	
WM 324		*Yarrowia lipolytica*	1945	USA	Environmental	Maize processing plant	KP132930	
WM 599	MMRL 210 B	*Yarrowia lipolytica*		USA	Clinical		EF568019.1	
WM 63	CBS 599	*Yarrowia lipolytica*	1921	Netherlands	Environmental	Rancid margarine	EF568021.1	
WM 900	CIN 02	*Yarrowia lipolytica*		New Zealand			OR196161	
WM 01.200	MMRL 237 A	*Yarrowia lipolytica*	1993	USA	Clinical		KP132919	
WM 02.460	CBS 5839 (T), H8.1, MH 44	*Zygoascus hellenicus*	1997	South Africa	Veterinary	Mastitic udder of cow	EF568100.1	

Appendix 2: List of Equipment

Equipment	Supplier
Automatic Pipettes (P2, P10, P200, P1000)	Pipetman; France
Bench Top Microfuge (5415C, 5415D)	Eppenford-Netherler-Hinz, Germany
Bench Top Centrifuge (5702G)	Eppenford-Netherler-Hinz, Germany
Canon Powershot G3 Photo Camera	Canon, Japan
Confocal Microscope	Nikon, Japan
Dry Block Heater	IKA Works, USA
DS-11 FX+ Spectrophotometer	DeNovix, USA
DS-11+ Spectrophotometer	DeNovix, USA
Electrophoresis Chamber	Bio-Rad Laboratories, Australia
Hemacytometer Cell Counter	Sigma-Aldrich, USA
HulaMixer	Thermo Fisher Scientific, USA
Magenetic Stirrer	IKA Works, USA
MiSeq System	Illumina, USA
MS1 Minishaker Vorex	IKA Works, USA
NanoDrop	Thermo Fisher Scientific, USA
Nanopore MinION Mk1B	Oxford Nanopore Technologies, United Kingdom
Promega Quantus™ Fluorometer	Promega, USA
Refrigerated Centrifuge 5910 Ri	Eppenford-Netherler-Hinz, Germany
Refrigerated Microcentrifuge	Sigma-Aldrich, USA
Qubit Flurometer	Promega, USA
Thermal Cycler (D-37085)	SensoQuest, Germany
UV Transilluminator TM40E (λ=302nm)	UniEquip, Germany

Appendix 3: List of Computer Software

Computer Software	Supplier
Albacore	Oxford Nanopore Technologies; Oxford, United Kingdom
Bedtools	University of Utah; Utah, USA
BUSCO	University of Geneva; Geneva, Switzerland
Bwa-mem	Wellcome Trust Sanger Institute; Cambridge, United Kingdom
DADA2	Stanford University; California, USA
Deepbiner	University of Melbourne; Melbourne, Australia
DnaSP	University of Barcelona; Catalonia, Spain
EMBOSS cons	European Bioinformatics Institute; Cambridge, United Kingdom
EPI2ME	Oxford Nanopore Technologies; Oxford, United Kingdom
Ete3 module	Centro De Biotechnologia Y Genomica De Plantas; Madrid, Spain
Microsoft Excel	Microsoft Corporation; Washington, USA
Flye	National Cancer Institute; Maryland, USA
Guppy	Google; California USA
IDBA_UD	University of Hong Kong; Hong Kong
Illumina bcl2fastq	Illumina; California, USA
Keras	Google; California, USA
Kraken2	Johns Hopkins University; Maryland, USA
Makeblastdb	Southern Methodist University; Texas, USA
Medaka	Oxford Nanopore Technologies; Oxford, United Kingdom
MEGA7	University of Tokyo; Tokyo, Japan
Minimap	Broad Institute; Massachusetts, USA
MinKNOW	Oxford Nanopore Technologies; Oxford, United Kingdom
Mothur	University of Massachusetts, Massachusetts, USA
MUSCLE	European Bioinformatics Institute; Cambridge, United Kingdom
NanoSim	Canada's Michael Smith Genome Science Centre; Vancouver, Canada
Nucleotide BLAST	National Institute of Health; Maryland, USA
NumPy	NumPy Steering Council
Porechop	Pennsylvania State University; Pennsylvania, USA
Python	Python Software Foundation; Delaware, USA
Python Pandas module	NumFOCUS; Texas, USA
Qcat	Oxford Nanopore Technologies; Oxford, United Kingdom
QIIME2	QIIME2 Development Team
Racon	Stanford University; California, USA
Samtools	Harvard University; Massachusetts, USA
Seqtk	Harvard University; Massachusetts, USA

Sequencher™	Gene Codes Corporation; Michigan, USA
Trimomatic	USADELLAB; Düsseldorf, Germany
What's In My Pot	Oxford Nanopore Technologies; Oxford, United Kingdom

Appendix 4: List of Reagents

Chemicals, Kits and Reagents	Supplier
2-Mercaptoethanol	Sigma-Aldrich, USA (cat. no. M6250-250ML)
Agarose	Sigma-Aldrich, USA (cat. no. A4718-100G)
Agencourt AMPure XP Beads	Beckman Coulter, USA (cat. no. A63881)
Buffer RLT	QIAGEN, Germany (cat. no. 79216)
Cetyl Trimethylammonium bromide (CTAB) buffer	Sigma-Aldrich, USA (cat. no. H6269)
Deoxynucleotides Set (dNTPs)	Meridian Bioscience, Australia (cat. no. BIO-39049)
DNA Maxi Blood Kit	QIAGEN, Germany (cat. no. 51192)
DNA Mini Blood Kit	QIAGEN, Germany (cat. no. 51104)
DNeasy Plant Mini Kit	QIAGEN, Germany (cat. no. 69106)
DNeasy PowerMax Soil Kit	QIAGEN, Germany (cat. no. 12988-10)
dsDNA Broad Range Assay	DeNovix, USA
Ethylenediaminetetraacetic acid (EDTA)	Sigma-Aldrich, USA (cat. no. 798681)
Ethanol	Sigma-Aldrich, USA (cat. no. E7023)
Ethidium Bromide	Sigma-Aldrich, USA (cat. no. 46067)
HL-SAN DNase	ArcticZymes Technologies, Norway
Isopropanol	Sigma-Aldrich, USA (cat. no. I9516)
Library Loading Bread Kit	Oxford Nanopore Technologies, United Kingdom (cat. no. EXP-LLB001)
Ligation Sequencing Kit	Oxford Nanopore Technologies, United Kingdom (cat. no. SQK-LSK108)
Ligation Sequencing Kit	Oxford Nanopore Technologies, United Kingdom (cat. no. SQK-LSK109)
Magnesium Chloride Solution (MgCl2)	Meridian Bioscience, Australia (cat. no. BIO-37026)
MyTaq™ DNA Polymerase	Meridian Bioscience, Australia (cat. no. BIO-21105)
Native Barcoding Expansion (PCR-free) Kit	Oxford Nanopore Technologies, United Kingdom (cat. no. EXP-NBD103)
NEB Blunt/TA Master Mix	New England Biolabs, USA (cat. no. M0367)
NEBNext Quick Ligation Reaction Buffer	New England Biolabs, USA (cat. no. M0367S)
NEBNext Ultra II End-Repair/ dA-tailing Module	New England Biolabs, USA (cat. no. E7546S)
NH4 PCR Buffer	Meridian Bioscience, Australia (cat. no. BIO-37025)
Nuclease Free Water	Sigma-Aldrich, USA (cat. no. W4502)
Phenol: Chloroform: Isoayl Alcohol	Sigma-Aldrich, USA (cat. no. P3803)
Phosphate Buffered Saline (PBS)	Sigma-Aldrich, USA (cat. no. P4417)
Primers and Oligonucleotides	Meridian Bioscience, Australia
Proteinase K	Sigma-Aldrich, USA (cat. no. P5568)
PureLink™ Genomic DNA Mini Kit	Thermo Fisher, USA (cat. no. K1820-01)
Q5 High-Fidelity DNA polymerase	New England Biolabs, USA (cat. no. M0515)
Quick T4 DNA Ligase	New England Biolabs, USA (cat. no. E6056)

Quick-DNA Fungal/Bacterial Miniprep Kit	Zymo Research, USA (cat. no. D6005)
RNAse T1	Thermo Fisher, USA (cat. no. EN0541)
Saponin	Sigma-Aldrich, USA (cat. no. 47036)
Sodium Chloride (NaCl)	Sigma-Aldrich, USA (cat. no. S3014)
Sodium dodecyl sulfate (SDS)	Sigma-Aldrich, USA (cat. no. 428015)
Sputum DNA Isolation Kit	Norgen, Canada (cat. no. 46200)
Tris-Hydrochloride (Tris-HCl)	Sigma-Aldrich, USA (cat. no. PHG0002)
Ultra-Deep Microbiome Prep10 Kit	Molzym, Germany
Zymo Research Genomic DNA Clean and Concentrator	Zymo Research, USA (cat. no. D4065)

Appendix 5: Long-read sequencing in fungal identification

This manuscript was published in Microbiology Australia on the 8th April 2022 (Volume 43, Issue 1, Pages 14-18). Hoang M.T.V. is the primary author and wrote and corrected the manuscript. Irinyi L. and Meyer W also wrote and corrected the manuscript and conceived the review. Meyer W. coordinated and supervised the review.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Minh Thuy Vi Hoang
30th June 2023

As a supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Professor Wieland Meyer
30th June 2023

CSIRO PUBLISHING

IN FOCUS
https://doi.org/10.1071/MA22006

Microbiology Australia

Long-read sequencing in fungal identification

Minh Thuy Vi Hoang[A,B], Laszlo Irinyi[A,B,C] and Wieland Meyer[A,B,C,D,E,*]

For full list of author affiliations and declarations see end of paper

***Correspondence to:**
Wieland Meyer
Molecular Mycology Research Laboratory, Centre for Infectious Diseases and Microbiology, Faculty of Medicine and Health, Sydney Medical School, Westmead Clinical School, The University of Sydney, Sydney, NSW 2006, Australia
Email: wieland.meyer@sydney.edu.au, wieland.meyer@curtin.edu.au

Received: 23 December 2021
Accepted: 11 February 2022
Published: 8 April 2022

Cite this:
Hoang MTV *et al.* (2022)
Microbiology Australia
43(1), 14–18. doi:10.1071/MA22006

© 2022 The Author(s) (or their employer(s)). Published by CSIRO Publishing on behalf of the ASM. This is an open access article distributed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND)

OPEN ACCESS

ABSTRACT

Long-read sequencing is currently supported by sequencing platforms from Pacific Biosciences and Oxford Nanopore Technologies, both of which generate ultra-long reads. Metabarcoding and metagenomics are the two approaches used when implementing sequencing. Metabarcoding involves the amplification and sequencing of selected nucleic acid regions, while in a metagenomic approach extracted nucleic acids are sequenced directly without prior amplification. Both approaches have associated advantages and disadvantages, which, in combination with long-read sequencing, provide a promising new approach for fungal identification and diagnosis of mycoses, on which we will reflect in this short review.

Keywords: diagnostics, DNA barcoding, fungal identification, long-read sequencing, metabarcoding, metagenomics, mycoses, next generation sequencing.

Long-read sequencing

Long-read sequencing technologies are characterised by the potential to generate ultra-long reads over 10 kb in one run.[1] Pacific Biosciences (PacBio) first released their long-read sequencing instrument in 2011 and most recently released the Sequel IIe (Fig. 1).[2] PacBio sequencing is based on the conversion of fluorescent signals produced when nucleotides are bound to a template strand via a polymerase. The template double-stranded DNA has two hairpin adaptors bound to each end and so sequencing continues around the template for the duration of the polymerase lifetime. PacBio sequencing has achieved high accuracy, which is reported to be 99.8%, and produces a high throughput with parallel sequencing of millions of template strands.[3] However, PacBio sequencing is a trade-off between read length and read quality: as longer template strands are sequenced, there is often a less accurate consensus sequence. Additionally, library preparation is estimated to be 1 day, which may not be suitable for time sensitive applications. Oxford Nanopore Technologies (ONT) began commercial release in 2015 with their MinION sequencer and has since released instruments and flow cells for a variety of throughput requirements (Fig. 2).[4] The basis of ONT sequencers are biological nanopores fixed in membranes, of which nucleotide strands travel through to cause a current change, which is then translated into a sequence. ONT sequencing read length is only limited by sample DNA length, quantity, and purity and as they provide a wide array of library preparation kits and throughput options it can be scaled to any potential use intended. Additionally, the low initial investment and small portable size of some ONT sequencers allow sequencing to be performed in locations outside the laboratory.[2] The major drawback of ONT is the lower read accuracy reported, although computation intensive bioinformatic tools are available to increase accuracy to 99%.[5] Samples that result in low DNA quantity, quality, and short read lengths do not take full advantage of ONT sequencing and impede sequencing by blocking and inactivating pores. These sequencing technologies are further discussed in other studies.[2]

Studies using PacBio and ONT long-read sequencers to sequence and identify fungal species using metabarcoding and metagenomic approaches are described below to demonstrate the potential use for long-read sequencers in fungal identification.

Metabarcoding

Metabarcoding involves the sequencing of targeted nucleic acid regions within environmental and clinical samples. This approach combines DNA barcoding with high throughput sequencing of specific taxonomic regions (barcodes). These barcodes, which have

Fig. 1. Pacific Biosciences Long Read Sequencing Platforms. The first commercial long read sequencer Sequel was released in 2015, supporting the SMRT Cell 1 M. The newer instruments Sequel II and Sequel IIe support the SMRT Cell 8 M, capable of simultaneous sequencing of 8 million template DNA strands. The Sequel IIe improves upon the previous iterations through software upgrades resulting in a reduction in analysis, file transfer and data storage requirements. Instrument pictures from the manufacturer's webpage (https://www.pacb.com/).

Fig. 2. Oxford Nanopore Long Read Sequencing Platforms. The flagship product, the MinION, is a palm-sized device supporting MinION flow cells containing 4 × 125 nanopore channels. GridION and PromethION are the higher throughput instruments, supporting 5 and 48 flow cells respectively. The Flongle flow cell contains 126 nanopore channels and is compatible with the MinION and GridION. A flow cell for the PromethION containing 2675 nanopore channels is also commercially available. Instrument pictures from the manufacturer's webpage (https://nanoporetech.com/).

high taxonomic coverage and high resolution, are amplified and sequenced (Fig. 3).[6] The generated barcode sequences are clustered into operational taxonomic units (OTUs) based on sequence similarity and compared against databases containing reference barcode sequences to provide an accurate representation of the microbial population. The advantages of this approach are the low level of genomic material required and faster and less complex computational analysis. However, as only specific regions are sequenced, information is limited to identification, although this may not be an issue for all studies. Prior amplification of a sample also potentially introduces bias, which may result in the inaccurate representation of the microbial community.[6] Previous microbial metabarcoding studies with short-read sequencing technologies used micro-barcodes that spanned less than 600 bp, which were often shorter than the full-length barcoding regions.[7] Long-read sequencers can span beyond the full-length barcode regions and resolve longer structural variations that are challenging for short-read sequencers, leading to higher discriminatory power.

Long-read fungal metabarcoding studies primarily use the full-length internal transcribed spacer (ITS) region of the rRNA gene cluster to identify fungal species, although shorter barcode regions are also used. To validate the ability of long-read sequencers to identify fungal species, ONT and PacBio have been used to identify members of mock fungal communities. The full-length ITS and the ITS1 regions were shown to accurately identify 16 and 26 fungal species within mock communities respectively using PacBio sequencing.[8,9] The full-length ITS region in conjunction with ONT sequencing has also successfully identified fungal species in mock communities.[10,11] The ITS region was found to be the superior locus for fungal identification in nanopore sequencing although, in a mock community with varying abundance of 16 fungal species, species level identification was only achieved for 1/3 of fungal species.[7] Metabarcoding of clinical samples with nanopore sequencing of the ITS region, has been explored extensively. Pathogens were identified from nine positive blood culture bottles which were then verified by routine diagnosis, one of which was a *Candida albicans* infection.[5] Type strains of five *Candida* species were also identified to sufficient (100–200×) coverage and nanopore sequencing errors did not affect correct species identification.[5] Full-length ITS nanopore sequencing has also identified potential pathogens in patient samples previously negative by traditional diagnostic methods.[11,12] PacBio has also been used in the clinical space to characterise the gut mycobiome of 14 healthy individuals with the ITS1 locus.[9] PacBio sequencing has primarily been applied to metabarcoding of environmental samples and has been demonstrated to outperform nanopore sequencing in such applications due to nanopore sequencing errors.[13] Metabarcoding of samples, such as tree roots,[14] soil,[13] mangrove sediments,[15] and lake water,[7] have revealed the potential of PacBio targeting the full-length ITS region to be used in broad ecological studies requiring accurate characterisation of the mycobiome of complex environmental samples. Metabarcoding of fungi using long-read sequencing has been established to be a promising avenue for the identification and characterisation of fungal species in clinical and environmental samples.

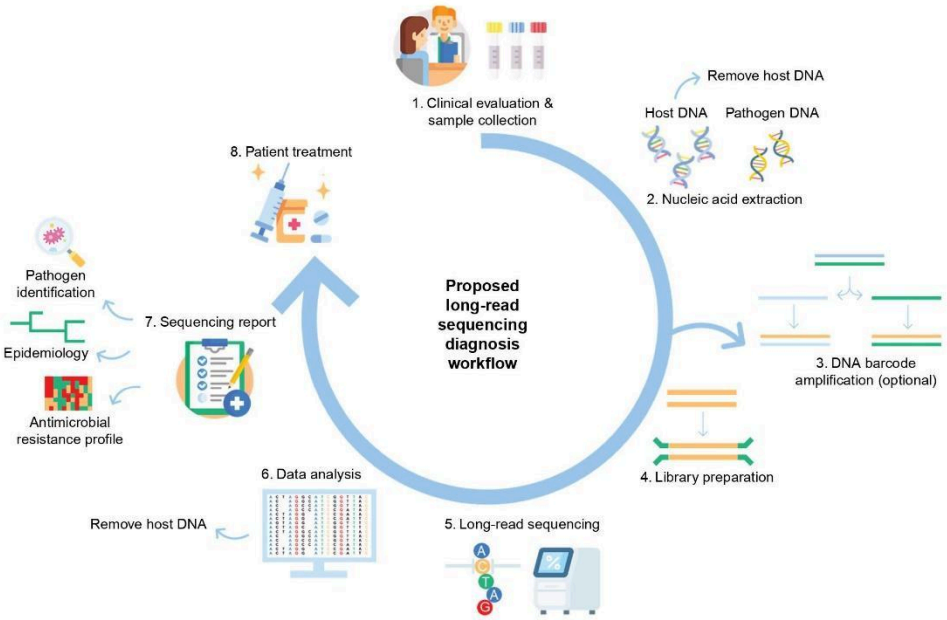


Fig. 3. Proposed Long-read Sequencing Diagnosis Workflow. Diagnosis of fungal infections would begin with clinical evaluation and patient sample collection. Nucleic acids would then be extracted from patient samples containing both host and pathogen DNA. Additional removal of host DNA may occur at this stage, and/or in metabarcoding studies. This is followed by DNA barcode amplification. In both metagenomic and metabarcoding studies, library preparation takes place before sequencing with long-read sequencers. Analysis of sequencing data would then be performed to generate a sequencing report which may include pathogen identification, epidemiological information, or an antimicrobial resistance marker (specific known mutations that confer resistance) maybe identified during metagenomic sequencing. This information in conjunction with the clinical evaluation then provides a diagnosis from which a treatment plan for the patient can be developed. Icons made by https://www.flaticon.com/authors/freepik.

Metagenomics

Metagenomic sequencing, also known as shotgun sequencing, aims to sequence all genetic material within a sample. In this approach, all genetic material (DNA or RNA) is extracted from the primary samples, which are then fragmented and undergo library preparation to suit the sequencing technology. The sequencing library then undergoes in-depth sequencing and data analysis (Fig. 3).[16] An advantage of metagenomics over metabarcoding is higher resolution, as more parts of the genomes of every organism in the sample are sequenced, potentially generating information beyond identification, such as antimicrobial resistance and virulence. The direct sequencing of genetic material eliminates the need for prior culturing and amplification, reducing the time from sample collection to identification, eliminating any bias that may occur due to these additional steps, resulting in a more accurate representation of the community composition.[17] In metagenomic sequencing studies, the overwhelming amount of background DNA compared to microbial DNA remains a challenge and methods to enrich microbial DNA have been developed.[18] Additionally, the cost and bioinformatic requirements of metagenomic sequencing is generally higher than those of metabarcoding studies. For full use of metagenomic sequencing reads, robust reference whole genome sequences are required. Long-reads generated in metagenomic studies are more correctly mapped to reference genomes and give high discriminatory power for accurate identification. However, the main current hurdles for long-read metagenomics based fungal identification are the fact that current genome databases lack adequate coverage for all fungal species.[19]

Metagenomic studies with long-read sequencers to identify fungal species are currently limited. However, preliminary studies have demonstrated their promising potential. PacBio sequencing of skin samples using a metagenomic approach identified a similar microbial community to short-read sequencing.[20] Additionally, metagenomic PacBio sequencing has been used in conjunction with short-read sequencing for genome assembly of fungal species in complex lichen samples.[21,22] Metagenomic nanopore sequencing identified pathogens from 87 patient samples in a single hospital study from a range of infections, including fungal infections, achieving sensitivity and specificity of 90.9 and 100% respectively, outperforming short-read sequencing.[23] The same approach was applied to three patient samples positive for *Pneumocystis jirovecii* and three negative respiratory samples.[19] All positive samples returned reads identified as

P. jirovecii. However, *P. jirovecii* was also detected in negative samples. Furthermore, fungal species were identified that are geographically restricted to areas that did not align with the patient's travel histories. These likely misidentifications were attributed to issues applying the bioinformatics tools to fungal identification. Additionally, all samples reported 77–95% *Homo sapiens* reads aside from one outlier (10%) demonstrating the high abundance of human background DNA.[19] Studies involving methods to overcome the limitations of metagenomic sequencing have emerged. The combination of metagenomic sequencing and whole genome amplification has been utilised to increase DNA quantity whilst maintaining the community composition to characterise the microbiome on the surface of oil paintings.[24] A method to deplete human DNA in clinical samples has been applied to respiratory samples and confirmed an *Aspergillus* infection previously diagnosed by culturing.[25] Additionally, this method also identified fungemia caused by *Candida glabrata* within 24 h whilst an extended culturing time (48–72 h to actionable results) was required for traditional identification.[25] Although there are currently drawbacks, with methods to overcome its limitations, long-read metagenomics is a favourable prospect for fungal identification.

Conclusion

Sequencing of genetic material has taken a significant step forward with the release of PacBio and ONT long-read sequencers. Their initial applications to fungal identification have indicated the promising potential for routine use in clinical diagnostics, microbiome characterisation, whole genome assembly and more, if the limiting factors of DNA extraction, low fungal-human sample DNA range, lack of reference sequences (DNA barcodes and whole genomes), and lack of bioinformatic tools, can be overcome. Implementation of long-read sequencing to diagnosis of fungal infections additionally requires a standardised workflow (Fig. 3). If successfully introduced into routine diagnosis of fungal infections it would drastically reduce turnaround time, from currently several days/weeks to less than 24 h (Fig. 3), enabling a timely and accurate induction of antifungal treatment, reducing mortalities, treatment and hospitalisation costs.

References

1. Pollard MO *et al.* (2018) Long reads: their purpose and place. *Hum Mol Genet* **27**, R234–R241. doi:10.1093/hmg/ddy177
2. Tedersoo L (2021) Perspectives and benefits of high-throughput long-read sequencing in microbial ecology. *Appl Environ Microbiol* **87**, e0062621. doi:10.1128/AEM.00626-21
3. Tedersoo L *et al.* (2018) PacBio metabarcoding of fungi and other eukaryotes: errors, biases and perspectives. *New Phytol* **217**, 1370–1385. doi:10.1111/nph.14776
4. Deamer D *et al.* (2016) Three decades of nanopore sequencing. *Nat Biotechnol* **34**, 518–524. doi:10.1038/nbt.3423
5. Ashikawa S *et al.* (2018) Rapid identification of pathogens from positive blood culture bottles with the MinION nanopore sequencer. *J Med Microbiol* **67**, 1589–1595. doi:10.1099/jmm.0.000855
6. Taberlet P *et al.* (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. *Mol Ecol* **21**, 2045–2050. doi:10.1111/j.1365-294X.2012.05470.x
7. Heeger F *et al.* (2018) Long-read DNA metabarcoding of ribosomal RNA in the analysis of fungi from aquatic environments. *Mol Ecol Resour* **18**, 1500–1514. doi:10.1111/1755-0998.12937
8. Mafune KK *et al.* (2020) A rapid approach to profiling diverse fungal communities using the MinION nanopore sequencer. *Biotechniques* **68**, 72–78. doi:10.2144/btn-2019-0072
9. Motooka D (2017) Fungal ITS1 deep-sequencing strategies to reconstruct the composition of a 26-species community and evaluation of the gut mycobiota of healthy Japanese individuals. *Front Microbiol* **8**, 238. doi:10.3389/fmicb.2017.00238
10. D'Andreano S *et al.* (2021) Rapid and real-time identification of fungi up to species level with long amplicon nanopore sequencing from clinical samples. *Biol Methods Protoc* **6**, bpaa026. doi:10.1093/biomethods/bpaa026
11. Wang M *et al.* (2020) Same-day simultaneous diagnosis of bacterial and fungal infections in clinical practice by nanopore targeted sequencing. *medRxiv* 2020.04.08.20057604. doi:10.1101/2020.04.08.20057604
12. Chan WS *et al.* (2020) Potential utility of targeted Nanopore sequencing for improving etiologic diagnosis of bacterial and fungal respiratory infection. *Diagn Pathol* **15**, 41. doi:10.1186/s13000-020-00960-w
13. Loit K *et al.* (2019) Relative performance of MinION (Oxford Nanopore Technologies) versus Sequel (Pacific Biosciences) third-generation sequencing instruments in identification of agricultural and forest fungal pathogens. *J Appl Environ Microbiol* **85**, e01368–19. doi:10.1128/AEM.01368-19
14. Marčiulyniene D *et al.* (2021) DNA-metabarcoding of belowground fungal communities in bare-root forest nurseries: focus on different tree species. *Microorganisms* **9**, 150. doi:10.3390/microorganisms9010150
15. Zhang ZF *et al.* (2021) High-level diversity of basal fungal lineages and the control of fungal community assembly by stochastic processes in mangrove sediments. *Appl Environ Microbiol* **87**, e0092821. doi:10.1128/AEM.00928-21
16. Thomas T *et al.* (2012) Metagenomics – a guide from sampling to data analysis. *Microb Inform Exp* **2**, 3. doi:10.1186/2042-5783-2-3
17. Laudadio I *et al.* (2018) Quantitative assessment of shotgun metagenomics and *16S* rDNA amplicon sequencing in the study of human gut microbiome. *OMICS* **22**, 248–254. doi:10.1089/omi.2018.0013
18. Charalampous T *et al.* (2019) Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. *Nat Biotechnol* **37**, 783–792. doi:10.1038/s41587-019-0156-5
19. Irinyi L *et al.* (2020) Long-read sequencing based clinical metagenomics for the detection and confirmation of *Pneumocystis jirovecii* directly from clinical specimens: a paradigm shift in mycological diagnostics. *Med Mycol* **58**, 650–660. doi:10.1093/mmy/myz109
20. Tsai YC *et al.* (2016) Resolving the complexity of human skin metagenomes using single-molecule sequencing. *mBio* **7**, e01948–15. doi:10.1128/mBio.01948-15
21. Greshake Tzovaras B *et al.* (2020) What is in *Umbilicaria pustulata*? A metagenomic approach to reconstruct the holo-genome of a lichen. *Genome Biol Evol* **12**, 309–324. doi:10.1093/gbe/evaa049
22. Wang B *et al.* (2021) Chromosome-scale genome assembly of *Fusarium oxysporum* strain Fo47, a fungal endophyte and biocontrol agent. *Mol Plant Microbe Interact* **33**, 1108–1111. doi:10.1094/MPMI-05-20-0116-A
23. Gu W *et al.* (2021) Rapid pathogen detection by metagenomic next-generation sequencing of infected body fluids. *Nat Med* **27**, 115–124. doi:10.1038/s41591-020-1105-z
24. Piñar G *et al.* (2020) Rapid diagnosis of biological colonization in cultural artefacts using the MinION nanopore sequencing technology. *Int Biodeterior Biodegradation* **148**, 104908. doi:10.1016/j.ibiod.2020.104908
25. Yang L *et al.* (2019) Metagenomic identification of severe pneumonia pathogens in mechanically-ventilated patients: a feasibility and clinical validity study. *Respir Res* **20**, 265. doi:10.1186/s12931-019-1218-4

Data availability. Data sharing is not applicable as no new data were generated or analysed during this study.

Conflicts of interest. The authors declare no conflicts of interest.

Declaration of funding. This study was supported by a National Health and Medical Research Council of Australia (NHMRC; grant no. APP1121936) to Wieland Meyer.

Author affiliations

[A]Molecular Mycology Research Laboratory, Centre for Infectious Diseases and Microbiology, Faculty of Medicine and Health, Sydney Medical School, Westmead Clinical School, The University of Sydney, Sydney, NSW 2006, Australia.

[B]Molecular Mycology Research Laboratory, Centre for Infectious Diseases and Microbiology, Westmead Institute for Medical Research, Westmead, NSW 2145, Australia.

[C]Sydney Institute for Infectious Diseases, The University of Sydney, Sydney, NSW 2006, Australia.

[D]Westmead Hospital (Research and Education Network), Westmead, NSW 2145, Australia.

[E]Curtin Medical School, Curtin University, Perth, WA 6102, Australia.

Biographies

Minh Thuy Vi Hoang is a PhD student at the University of Sydney at the Sydney Medical School. Her doctoral work explores the application of long-read sequencers to the diagnosis of fungal diseases. Prior, she received a Bachelor of Medical Science (First Class Honours) in Microbiology at the University of Sydney.

Dr Laszlo Irinyi is a Postdoctoral Fellow at Westmead Institute for Medical Research, Westmead, NSW, Australia and his research focuses on the identification and molecular taxonomy of human and animal pathogenic fungi. Currently his focus is on the adaptation of new generation sequencing technologies in routine diagnostics for fungal identification.

Professor Wieland Meyer is a Molecular Medical Mycologist and academic at the Faculty of Medicine and Health, Sydney Medical School; The University of Sydney, Associate Dean of Curtin Medical School, Curtin University, and the Fundação Oswaldo Cruz (FIOCRUZ) in Rio de Janeiro, Brazil, heading the MMRL within the CIDM, Westmead Institute for Medical Research, with a PhD in fungal genetics from the Humboldt University of Berlin, Germany. His research focuses on phylogeny, molecular identification, population genetics, molecular epidemiology, and virulence mechanisms of human and animal pathogenic fungi. He is the Convener of the Mycology Interest Group of ASM, and the President of the International Mycological Association (IMA).