Nanopore ITS Sequencing for Fungal Species Identification in Clinical Isolates

HM. Adriaanse 1

1 Avans University of Applied Sciences, Academy of Life Sciences and Technology, Bioinformatics

In collaboration with the Institute of Neuropathology in Münster, Germany

Corresponding author: [hm.adriaanse@student.avans.nl](mailto:hm.adriaanse@student.avans.nl)

Supervisor: Eefje Schrauwen

Date: 20-6-2025

**ABSTRACT**

Accurate identification of fungal pathogens is essential in clinical settings, especially when traditional culturing methods fail. This study evaluated whether Nanopore sequencing of PCR-amplified ITS regions can support species-level identification of clinical fungal isolates. Ten isolates were amplified using the ITS1-F\_KYO2a and RCA95m primers and sequenced with Oxford Nanopore technology. Three bioinformatics approaches were compared: (1) taxonomic classification using GermGenie, (2) mapping-based specificity assessment using Minimap2 and Samtools, and (3) identification from consensus sequences generated by Flye or wf-amplicon, followed by ITSx and BLAST.

The wf-amplicon variant calling workflow correctly identified 9 out of 10 isolates and outperformed both Flye and wf-amplicon de novo. GermGenie allowed rapid screening but misclassified most isolates, likely due to database limitations and off-target amplification. Mapping results showed large differences in specificity, influenced by genomic divergence and primer binding. However, identification success did not consistently correlate with mapping percentage or read quality. This highlights the importance of workflow design and error correction in ITS-based fungal classification.

Although this study used high-quality DNA, the methods and findings are relevant for future work on degraded samples, such as FFPE tissue. Improved performance may be achieved by using well-annotated ITS databases and including multiple reference genomes per species. The pipeline presented here provides a flexible approach for fungal identification using long-read sequencing.

Keywords: Nanopore sequencing, ITS region, PCR specificity, taxonomic classification

Index

[INTRODUCTION 3](#_Toc201242189)

[THEORETICAL BACKGROUND 3](#_Toc201242190)

[**Clinical Importance of Fungal Infections** 3](#_Toc201242191)

[**The ITS Region as a Fungal DNA Barcode** 4](#_Toc201242192)

[**ITS Sequencing from FFPE Tissue: Challenges and Strategies** 4](#_Toc201242193)

[**Next-Generation Sequencing for Fungal Detection** 4](#_Toc201242194)

[**Primer Design for Broad Fungal Amplification** 4](#_Toc201242195)

[**Challenges in Fungal Taxonomic Classification** 5](#_Toc201242196)

[**Quality Control in Nanopore Sequencing** 5](#_Toc201242197)

[**ITSx for Targeted Extraction of ITS Regions** 5](#_Toc201242198)

[**GermGenie and Reference Databases for ITS Classification** 5](#_Toc201242199)

[**Read Mapping with Minimap2 and Bedtools** 5](#_Toc201242200)

[MATERIAL AND METHODS 6](#_Toc201242201)

[**Materials** 6](#_Toc201242202)

[**Methods** 6](#_Toc201242203)

[RESULTS 7](#_Toc201242204)

[**Read quality and length assessment using NanoPlot and Filtlong** 7](#_Toc201242205)

[**Taxonomic classification using GermGenie** 8](#_Toc201242206)

[**Accuracy of Species Identification Based on Consensus Sequences** 9](#_Toc201242207)

[**Association Between Primer Recovery, Read Mapping, Abundance, and Tool Performance** 10](#_Toc201242208)

[DISCUSSION 11](#_Toc201242209)

[CONCLUSION 12](#_Toc201242210)

[AVAILABILITY OF DATA AND MATERIALS 12](#_Toc201242211)

[ACKNOWLEDGMENTS 12](#_Toc201242212)

# INTRODUCTION

Infectious diseases continue to pose a major challenge in clinical diagnostics, especially in immunocompromised patients where rapid and accurate pathogen identification is essential for guiding treatment. Traditional diagnostic methods, such as microbial culturing, are often slow and unreliable for organisms that are difficult or impossible to cultivate. Fungal pathogens, in particular, are notoriously challenging to culture and identify using conventional approaches. As a result, molecular diagnostics based on sequencing technologies are increasingly being adopted for direct detection of microbial DNA from clinical samples [1].

Among available genetic markers, the internal transcribed spacer (ITS) region of the fungal rRNA operon has been widely accepted as the formal DNA barcode for fungi, owing to its high interspecific variability and universal presence in fungal genomes [2]. Amplification and sequencing of this region allow for species-level identification in most fungal lineages. However, the full ITS region, which may span up to 2.5 to 3 kbp including flanking rRNA genes, is often difficult to amplify in degraded DNA samples such as formalin-fixed paraffin-embedded (FFPE) tissue. In bacterial diagnostics similar challenges have been overcome by targeting shorter overlapping 16S rRNA gene fragments that enable accurate taxonomic classification even in highly fragmented DNA. [3]. In contrast, a validated fragment-based approach for fungal ITS sequencing remains underexplored.

This study was initiated in order to investigate whether Nanopore sequencing of PCR-amplified ITS regions, using the primers ITS1-F\_KYO2a and RCA95m, permits accurate fungal species identification from clinical fungal isolates. This approach serves as a model for evaluating the potential of ITS-based diagnostics in settings with degraded DNA, such as FFPE tissues, even though the current study focuses on DNA extracted from high-quality clinical fungal isolates.

To achieve this, ten clinically relevant fungal isolates were selected, and a large genomic region spanning ITS1, 5.8S, and ITS2 was amplified using the primers ITS1-F\_KYO2a and RCA95m. The resulting amplicons were sequenced using Oxford Nanopore Technology (ONT), generating long-read data in FASTQ format. Initial quality control of the raw Nanopore reads was performed using NanoPlot and Filtlong to assess sequencing metrics and retain high-quality reads based on length and Phred score thresholds. These quality-filtered FASTQ reads were subsequently used in two parallel analyses. First, reads were aligned to fungal reference genomes using Minimap2 to assess mapping specificity. To evaluate potential off-target amplification, overlaps between mapped reads were examined using bedtools. Second, the filtered reads were subjected to direct taxonomic classification using GermGenie, which was applied to the FASTQ data in combination with a curated fungal ITS reference database. This allowed assessment of whether accurate species-level identification could be achieved without the need for prior consensus generation. In a separate pipeline, consensus sequences were generated from the filtered reads using Flye and Wf-amplicon. These consensus sequences were then processed with ITSx, to identify and extract the ITS1 and ITS2 subregions. The extracted ITS regions were then analyzed using BLASTn searches against the UNITE fungal ITS reference database, enabling accurate taxonomic assignment based on sequence similarity.

# THEORETICAL BACKGROUND

## **Clinical Importance of Fungal Infections**

Fungal infections are increasingly recognized as an important cause of illness and death, particularly in individuals with weakened immune systems. This includes patients undergoing chemotherapy, organ transplantation or long-term treatment with specific medications. Invasive fungal infections of the central nervous system (CNS), although relatively uncommon, are often linked to poor outcomes and high mortality. Early and accurate detection of the fungal pathogen is crucial to ensure timely and effective antifungal treatment. Traditional diagnostic methods, such as histological examination and fungal culture, are often slow and lack sensitivity. They also frequently fail to identify the fungus at the species level, especially when dealing with slow-growing, non-culturable or visually similar organisms. Molecular techniques such as PCR and sequencing have emerged as faster and more reliable alternatives, offering improved accuracy and speed in fungal identification. [4]

## **The ITS Region as a Fungal DNA Barcode**

The internal transcribed spacer (ITS) region of the nuclear ribosomal DNA (rDNA) is considered the universal barcode for fungi due to its presence across all fungal taxa and its high interspecific variability. It comprises two hypervariable regions, ITS1 and ITS2, separated by the conserved 5.8S rRNA gene. ITS1 lies between the 18S and 5.8S rRNA genes, and ITS2 between the 5.8S and 28S genes. This architecture enables both the design of universal primers and the discrimination between closely related species. ITS sequencing has become the gold standard for fungal taxonomic identification in both environmental and clinical contexts. It enables the detection of diverse fungi, including rare and unculturable species. This can be directly from complex samples such as tissue biopsies or cerebrospinal fluid. Large publicly available databases such as UNITE and GenBank contain thousands of ITS reference sequences, facilitating species-level identification based on sequence homology. However, challenges remain. Intragenomic ITS variability within a single fungal genome can lead to unclear results, and closely related species may have nearly identical ITS sequences. [5] [6]

## **ITS Sequencing from FFPE Tissue: Challenges and Strategies**

Formalin-fixed, paraffin-embedded (FFPE) tissue samples represent a valuable resource in clinical and retrospective research, but the fixation process introduces significant challenges for molecular analyses. Formalin-induced crosslinking and fragmentation severely compromise DNA quality, often reducing fragment lengths to 100–300 base pairs and introducing chemical modifications such as deamination or base substitutions [7] [8]. These changes hinder PCR amplification, particularly of longer regions like full-length ITS1-5.8S-ITS2. Amplification artifacts, low yields, and sequencing errors are common in FFPE-derived DNA. To overcome this, several strategies have been explored: using high-fidelity polymerases resistant to inhibitors, targeting shorter ITS subregions to accommodate fragmented DNA, or applying hybrid capture enrichment methods. Enzymatic repair protocols like uracil-DNA glycosylase (UDG) treatment have also been tested, but none fully resolve the issues of DNA fragmentation and crosslinking [9] [10].

## **Next-Generation Sequencing for Fungal Detection**

Next-Generation Sequencing (NGS) provides a high-throughput, culture-independent means of detecting and identifying pathogens in clinical samples. In amplicon-based NGS, specific genetic loci are amplified by PCR prior to sequencing. For fungal identification, the ITS region is most commonly targeted due to its taxonomic resolution. NGS is especially useful for infections with multiple microbes or when only small amounts of DNA are present. It can detect fungal DNA even if there is a lot of human or bacterial DNA in the sample. Also, it doesn’t require the fungi to be alive, which is important for formalin-fixed samples. Oxford Nanopore Technology (ONT), which produces long sequencing reads in real time, offers unique advantages for spanning entire ITS regions and resolving complex mixtures, albeit at the cost of higher per-read error rates. [3] [6]

## **Primer Design for Broad Fungal Amplification**

Effective primer selection is essential for the reliable amplification of the ITS region across a wide range of fungal taxa. In this study, the forward primer ITS1-F\_KYO2a (Fw1) and the reverse primer RCA95m (Rv5) was used. ITS1-F\_KYO2a binds near the 3′ end of the small subunit (SSU) rRNA gene, directly upstream of the ITS1 region and the 5.8S rRNA gene. The RCA95m primer binds just after the ITS2 region and extends into the large subunit (LSU) rRNA gene, allowing amplification of a longer ITS region that includes ITS1, 5.8S, and ITS2. This primer combination is particularly suitable for long-read sequencing platforms such as Oxford Nanopore, which benefit from longer input fragments for improved base calling and taxonomic resolution. Amplifying the full ITS1–5.8S–ITS2 region provides more phylogenetic information than partial ITS1 or ITS2 fragments alone, thereby enhancing the accuracy of fungal identification. A schematic overview of the fungal rRNA operon and the relative positions of the primers used in this study is shown in Figure 1.

Figure 1: Schematic representation of the fungal rRNA operon, highlighting the primers used in this study for ITS region amplification. The forward primer ITS1-F\_KYO2a (Fw1) targets the ITS1 region between the small subunit (SSU) and the 5.8S rRNA gene, while the reverse primer RCA95M (Rv5) extends beyond ITS2 into the large subunit (LSU) rRNA gene.

Afbeelding met tekst, schermopname, Lettertype, nummer

Automatisch gegenereerde beschrijving

## **Challenges in Fungal Taxonomic Classification**

Fungal taxonomy remains challenging due to intragenomic variation, high genetic diversity, and incomplete reference databases. Unlike the 16S rRNA gene used in bacterial classification, fungal markers such as ITS can exhibit multiple copies within a genome that are not identical, leading to intra-individual sequence variation [11]. Misannotated or missing entries in public databases further complicate classification, particularly for rare or poorly studied fungi. Metagenomic samples such as FFPE tissue, may also contain contaminant DNA that generates spurious identifications. Better reference databases and using more than one genetic marker are important for accurate fungal classification. [12]

## **Quality Control in Nanopore Sequencing**

Given the relatively high error rates of nanopore sequencing, careful quality control is essential. Tools such as NanoPlot and Filtlong are employed to assess and filter raw reads prior to downstream analysis. NanoPlot provides graphical summaries of read length, quality, and base composition, enabling visual assessment of sequencing performance. Filtlong filters reads based on quality scores and length thresholds, retaining only high-confidence sequences for analysis. This step ensures the quality of downstream processes, such as taxonomic classification or alignment. [13] [14]

## **ITSx for Targeted Extraction of ITS Regions**

ITSx is a specialized tool for extracting ITS1, ITS2, and 5.8S rRNA regions from raw DNA sequences. It uses Hidden Markov Models (HMMs) trained on fungal rDNA to accurately identify and separate ITS regions from flanking sequences. By isolating only the relevant barcode regions, ITSx improves both taxonomic resolution and computational efficiency. This is particularly important for long-read sequencing data, where reads may span entire operons or include irrelevant non-target regions. ITSx requires FASTA-formatted input, but sequencing data is typically generated in FASTQ format. To generate FASTA input for ITSx, tools such as Flye (for de novo assembly), Wf-amplicon (for consensus sequence generation), or Seqtk (for direct FASTQ-to-FASTA conversion) can be used. [15]

## **GermGenie and Reference Databases for ITS Classification**

GermGenie is a taxonomic classifier built on the Emu framework, which was originally developed for bacterial 16S rRNA analysis. In this study, GermGenie was adapted to classify fungal ITS reads by replacing the standard 16S database with a ITS reference database such as UNITE. Like EMU, it matches reads to the reference and produces taxon-level abundance profiles. The accuracy of species identification depends strongly on the quality and completeness of the reference database used. [16]

## **Read Mapping with Minimap2 and Bedtools**

To validate amplification success and assess mapping specificity, reads can be aligned to fungal reference genomes using Minimap2. This is a long-read aligner optimized for platforms like ONT. Minimap2 uses a seed-chaining algorithm to align even error-prone or fragmented reads with high speed and accuracy. In this study, aligned reads were further processed using Bedtools to determine whether they mapped specifically or aspecifically to off-target regions. This post-mapping validation is essential for evaluating PCR specificity and detecting potential mispriming events. [17] [18]

# MATERIAL AND METHODS

## **Materials**

Sequencing data were obtained from ten clinical fungal isolates. DNA was extracted and PCR-amplified using primers ITS1-F\_KYO2a (forward) and RCA95m (reverse), targeting the ITS1–5.8S–ITS2 region (~2,600–3,000 bp). Oxford Nanopore MinION sequencing produced single-end reads in FASTQ format. Raw files per barcode (barcode01–barcode10) were stored in the project\_data/ directory.

Taxonomic identification was performed using BLAST against the UNITE fungal ITS database (UNITE\_public\_19.02.2025.fasta.gz, doi:10.15156/BIO/3301227) and GermGenie classification against a custom EMU ITS database (sh\_general\_release\_10.05.2021.tgz, doi:10.15156/BIO/1280049).

Reference genomes of ten fungal species (e.g*., Candida albicans*, *Lichtheimia ramosa*, *Aspergillus fumigatus*) were retrieved from NCBI RefSeq and stored in the All\_fna/ directory for mapping-based validation.

These materials are described in more detail in the Materials section of the project repository (<https://github.com/DaanAdriaanse/fungal-ITS-identification-pipeline#materials>).

## **Methods**

**Process Description**

The full analysis was divided into a preprocessing phase and three subworkflows. Each step is described in detail in the GitHub repository (<https://github.com/DaanAdriaanse/fungal-ITS-identification-pipeline>).

**Preprocessing**

Raw reads were quality filtered using Filtlong and evaluated using NanoPlot. This step is documented in the Preprocessing section of the repository (<https://github.com/DaanAdriaanse/fungal-ITS-identification-pipeline#preprocessing-read-filtering--quality-control>).

**Sub-Workflow 1: Mapping and Specificity**

This workflow assessed whether reads mapped to the intended ITS region of their species: Reads were aligned to reference genomes using Minimap2. Mapping output was processed with Samtools and analyzed with Bedtools and BLASTn was used to locate primer binding sites to calculate the percentage of specific mapped reads. This workflow is described in Sub-Workflow 1 (<https://github.com/DaanAdriaanse/fungal-ITS-identification-pipeline#sub-workflow-1-mapping--specificity>).

**Sub-Workflow 2: GermGenie Classification**

Unassembled filtered reads were directly classified taxonomically using GermGenie, powered by the EMU classifier and a prebuilt UNITE ITS database. This rapid screening method is explained in Sub-Workflow 2 (<https://github.com/DaanAdriaanse/fungal-ITS-identification-pipeline#sub-workflow-2-germgenie>).

**Sub-Workflow 3: Consensus-Based/Convertfasta-based Identification**

Reads were converted or assembled into consensus sequences and processed via: Flye (de novo assembly), wf-amplicon (variant calling and de novo mode, seqtk (FASTQ to FASTA conversion). The ITS regions were extracted using ITSx, then identified via BLASTn using the UNITE+INSD database. See Sub-Workflow 3 for details (<https://github.com/DaanAdriaanse/fungal-ITS-identification-pipeline#sub-workflow-3-consensus-basedconvertfasta-based-identification>).

**Statistical Analysis**

This project focused on the implementation and validation of a bioinformatics pipeline for fungal ITS identification. As no hypothesis testing or experimental comparison between treatment groups was performed, traditional statistical analysis and power calculations were not applicable. Instead, performance and accuracy were evaluated descriptively using per-sample read-based values. These included the percentage of reads mapping to the expected ITS region (mapping specificity), the relative abundance of taxa detected by EMU classification, and BLAST-based identification values. Results were interpreted per barcode to assess the precision, specificity and overall reliability of each subworkflow in the pipeline.

**Programming and Script Availability**

All scripts and environments used in this project are fully documented in the GitHub repository. Each preprocessing step and subworkflow includes its own command-line scripts, Conda installation instructions, and organized input/output folder structure. This setup allows the pipeline to be run step-by-step or adapted for future projects.

**Implementation**

This project involved the development of a modular command-line pipeline for fungal ITS identification based on Nanopore amplicon data. The pipeline was structured in separate parts that could be executed independently or as a full workflow. The main components included preprocessing for read filtering and quality control, mapping to fungal reference genomes, and downstream validation of mapped regions. Mapping results were processed further using Samtools to sort, index and extract aligned reads. Bedtools was then used to convert BAM files to BED format and to merge or summarize mapped genomic regions. Taxonomic classification was carried out using GermGenie, which applies the EMU classifier with a prebuilt UNITE ITS database. Also a consensus-based identification workflow was set up using Flye and wf-amplicon, both in de novo and reference-guided modes. All FASTA sequences, whether generated by assembly or direct conversion, were processed using ITSx to extract ITS1 and ITS2 regions. These were then identified through BLASTn searches against the UNITE+INSD database.

# RESULTS

## **Read quality and length assessment using NanoPlot and Filtlong**

Nanopore sequencing was performed on PCR-amplified ITS regions from ten clinical fungal isolates. The raw sequencing output (FASTQ files) was first analyzed using NanoPlot to visualize the distribution of read length and Phred quality scores per sample. This initial quality check revealed considerable variation in both parameters across the datasets This filtering step was performed to eliminate low-quality or short reads that could compromise taxonomic classification, while still maintaining sufficient data coverage per isolate.

Table 1: The mean read length, average Phred quality score, and total number of reads per isolate, both before and after Filtlong filtering.

Afbeelding met tekst, schermopname, nummer, Lettertype

Door AI gegenereerde inhoud is mogelijk onjuist.

Mean read lengths after filtering varied between 1,910 and 4,360 bp, confirming the success of long-range PCR in most samples. *C. albicans* and *C. neoformans* showed the longest filtered reads (>3900 bp), reflecting efficient amplification and high-quality template DNA. *A. fumigatus* and *A. flavus* had shorter average read lengths, yet their filtered datasets still met the quality requirements for ITS identification. The isolate *T. indotineae* yielded the highest total read count, both before (39,579 reads) and after filtering (22,343 reads), and also achieved the highest mean Phred score (21.0), indicating excellence sequencing performance.

these results show that all filtered datasets had adequate read quality and length to support accurate identification of the ITS region and consistent species-level classification. This highlights the reliability of the sequencing and filtering workflow across a range of clinically relevant fungal isolates

## **Taxonomic classification using GermGenie**

Following quality control, taxonomic classification was performed using a modified version of GermGenie, adapted for fungal ITS sequence data. The top four most abundant fungal species per sample, as classified by GermGenie are visualized in Figure 4. These results demonstrate that the tool was able to differentiate a variety of fungal taxa across the ten clinical isolates; however, its accuracy in species-level identification was inconsistent.

Afbeelding met tekst, schermopname, diagram, Kleurrijkheid

Door AI gegenereerde inhoud is mogelijk onjuist.Figure 2: Relative abundance of the top four fungal species detected per clinical sample based on GermGenie ITS classification. Each bar represents one isolate, with stacked colors indicating the four most abundant taxa.

Table 2: Relative abundance of the fungal species, as determined by GermGenie.

Afbeelding met tekst, schermopname, Lettertype, nummer

Door AI gegenereerde inhoud is mogelijk onjuist. Of the ten samples, only one isolate, *M. pachydermatis*, was correctly classified with high confidence, showing a dominant relative abundance of 86.62 percent, in agreement with the expected species (table 4). For all other samples, GermGenie failed to identify the correct dominant species. For example, *T. indotineae* was not detected at all (0 percent), despite being one of the isolates with the highest number of high-quality reads. Similarly, *L. ramosa*, *A. fumigatus*, *C. tropicalis*, and *C. parapsilosis* were either absent or detected at low abundance in samples where they were known to be present. Unexpected taxa were frequently reported instead, including species from *C. Albicans, Serendipita spp*., and *Saccharomycetales*. The presence of these taxa could not be biologically confirmed and may reflect misclassification, database bias, or spurious sequence matches due to limited resolution in the classifier.

These findings suggest that GermGenie in its current fungal configuration, lacks the specificity and taxonomic precision required for reliable ITS-based fungal identification across diverse clinical samples. Further development, including improved database curation and fungal-specific classification models, is necessary before the tool can be considered suitable.

## **Accuracy of Species Identification Based on Consensus Sequences**

The performance of three consensus generation tools was evaluated: Flye, wf-amplicon de novo assembly, and wf-amplicon variant calling. Each tool was tested on ten clinical fungal isolates, and results were assessed based on three criteria: successful generation of a consensus sequence, detection of ITS1 and ITS2 regions using ITSx, and whether the top BLAST hit matched the expected species.

Table 3: Consensus generation, ITS region detection, and species-level identification for ten clinical fungal isolates using three reconstruction tools: Flye, wf-amplicon de novo, and wf-amplicon variant calling. For each tool, the presence of a consensus sequence, the detection of ITS1 and ITS2 regions (via ITSx), and the top BLAST hit are reported.

Afbeelding met tekst, schermopname, nummer, Lettertype

Door AI gegenereerde inhoud is mogelijk onjuist.

Flye generated consensus sequences in five out of ten isolates. The consensus lengths varied from 1,544 base pairs in *C. neoformans* to 3,523 base pairs in *M. pachydermatis*. ITSx successfully detected both ITS1 and ITS2 regions in only four of these five cases. However despite ITS detection, only two samples (*C. neoformans* and *L. ramosa*) were correctly identified by BLAST. The remaining three produced incorrect matches. *A. fumigatus* and *A. flavus* were both misidentified as *C. parapsilosis*. In *T. indotineae* the top BLAST hit was *Malassezia sp.*, a taxonomically unrelated organism. In the other five samples, Flye failed to produce a usable consensus and no ITS regions or BLAST results were available. These findings suggest that Flye often fails to reconstruct usable ITS sequences and lacks reliability for ITS-based fungal identification.

Wf-amplicon de novo generated consensus sequences in four isolates, with lengths ranging from 2,775 to 3,523 base pairs. ITS1 and ITS2 regions were detected in three out of these four cases, although in *A. fumigatus* no ITS regions were identified by ITSx. Only *M. pachydermatis* was correctly classified by BLAST. In *T. indotineae* ITS regions were present (694 bp and 274 bp), but the BLAST result was again incorrect (*Malassezia sp*.). No consensus could be produced for the majority of the samples. These results demonstrate that wf-amplicon de novo performs inconsistently and cannot be considered a reliable approach for routine ITS-based identification.

Wf-amplicon variant calling generated consensus sequences for all ten isolates. The consensus lengths ranged from 492 to 797 base pairs. ITSx detected both ITS1 and ITS2 regions in every case, with ITS1 lengths ranging from 123 to 338 base pairs, and ITS2 lengths ranging from 139 to 342 base pairs. BLAST analysis returned the correct species for nine out of ten isolates. Accurate identifications were obtained for *C. albicans*, *C. neoformans*, *L. ramosa*, *T. asahii*, *T. indotineae*, *A. fumigatus*, *A. flavus*, *C. parapsilosis*, and *M. pachydermatis*. Only *C. tropicalis* was misidentified as *Oligophagomyces castellii*. The consistent success across all isolates, coupled with complete ITS detection and high species-level accuracy, indicates that wf-amplicon variant calling is the most reliable method for ITS-based fungal identification among the tested approaches.

## **Association Between Primer Recovery, Read Mapping, Abundance, and Tool Performance**

To examine whether experimental parameters affected identification success, we compared the number of BLAST-matched primers, the percentage of specifically mapped reads, relative abundance (GermGenie) and the identification performance of Flye, Wf-amplicon de novo, and Wf-amplicon variant calling across ten fungal isolates (Table 4).

Table 4: shows the number of BLAST-matched forward and reverse primers per isolate, the percentage of specifically mapped reads, and the relative abundance as estimated by GermGenie. Identification outcomes are listed for Flye, wf-amplicon de novo, and wf-amplicon variant calling, with ✓ indicating a correct species-level match and - indicating a misclassification or no identification.

Afbeelding met tekst, schermopname, nummer, Lettertype

Door AI gegenereerde inhoud is mogelijk onjuist.

The number of matched primers ranged from 1 to 66 per direction. In some cases, isolates with many primer matches and high mapping rates were correctly identified. *L. ramosa* showed 66 forward and 66 reverse matches, a mapping rate of 98.81%, and was correctly classified by Flye and the variant workflow. Similarly, *M. pachydermatis* had 99.98% mapping, high relative abundance (86.62%), and was correctly identified by Wf-amplicon de novo and variant calling.

However, such cases did not reflect a consistent trend. *C. albicans*, with only one forward and one reverse primer, 100% mapping, and 11.86% relative abundance, was not correctly identified by Flye or de novo. *C. parapsilosis*, with similarly favorable mapping and primer recovery, was also misclassified by both tools but correctly identified by the variant workflow. These examples show that correct identification was not reliably predicted by high mapping rates or low primer counts alone.

Extremely low mapping percentages (<10%) were observed in *T. indotineae* (1.08%) and *A. fumigatus* (8.52%), neither of which were identified correctly by Flye or Wf-amplicon de novo. Despite this, both were successfully resolved by the variant workflow. Other isolates with moderate mapping values, such as *A. flavus* (56.01%) and *T. asahii* (58.17%), were also only correctly classified by Wf-amplicon variant calling.

Relative abundance values from GermGenie did not show a consistent relationship with classification performance. While *M. pachydermatis* had the highest abundance and was correctly identified, *C. albicans* and *C. neoformans* also had relatively high abundance (>11%) but were not consistently classified across tools. In contrast, *T. asahii* and *A. flavus*, which had low relative abundances (2.37% and 0.12%, respectively) were successfully identified using the variant method.

Flye correctly identified 2 out of 10 samples, Wf-amplicon de novo identified 1, and Wf-amplicon variant calling identified 9. No direct or consistent relationship was observed between the number of BLAST-matched primers, specific mapping percentage, or abundance and the success of identification across all tools. Only the Wf-amplicon variant workflow consistently produced correct classifications, even in samples with limited primer recovery, low mapping, or low abundance.

# DISCUSSION

This study evaluated the applicability of using Nanopore-based sequencing of the fungal ITS region for species-level identification of clinical isolates, with the broader goal of informing workflows applicable to FFPE-derived DNA. All analyses were based on the same input: quality-filtered Nanopore FASTQ reads of PCR-amplified ITS regions from ten fungal isolates. The study workflow was divided into three parallel approaches: taxonomic screening using GermGenie, specificity analysis using Minimap2 and Samtools, and ITS identification using reconstruction tools Flye, Wf-amplicon de novo, and Wf-amplicon variant calling in combination with ITSx and BLAST.

The main finding was that the Wf-amplicon variant calling workflow produced the highest accuracy, correctly identifying 9 out of 10 clinical isolates at species level. Flye only produced usable sequences for five samples, of which only two were correctly classified. Similarly, wf-amplicon de novo generated results for four isolates, but only one was correctly identified. These differences highlight the reliability of the variant workflow in clinical ITS sequencing. These differences in performance can be explained by how each workflow processes the data. The variant calling mode aligns reads directly to a reference sequence and uses Medaka to call variants and generate a polished consensus. This guided approach helps preserve sequence accuracy even when the raw reads contain errors and is well suited for clinical samples with degraded or noisy input [19]. The de novo mode assembles a draft consensus without a reference, using miniasm or spoa for shorter amplicons, and then polishes the result with Medaka. While this method can reconstruct full-length amplicons, it is more sensitive to low coverage and sequence variation, which may lead to errors or loss of sequences. [20] Flye builds assemblies entirely from scratch without a reference. This makes it more prone to fragmented or misassembled ITS sequences, especially when coverage is low or when repetitive regions interfere with assembly [21].

GermGenie was applied directly to the filtered FASTQ reads to assess whether rapid reference-based identification could be achieved without reconstruction. Using a fungal ITS reference database, GermGenie correctly identified *M. pachydermatis* but misclassified most other isolates and often detected unrelated taxa across samples. These results show that while GermGenie enables fast screening, its reliability for ITS-based fungal classification remains limited without improved database quality. This may be due to an incomplete or poorly curated reference database and possible off-target amplification during PCR, which can introduce non-specific reads and lower taxonomic resolution.

To assess PCR specificity, Minimap2 and Samtools were used to align filtered reads against fungal reference genomes. Specific mapping percentages ranged from 1.08% (*T. indotineae*) to 100% (*C. albicans*). The observed variation in mapping percentages across samples can be attributed primarily to two factors: genomic divergence and primer specificity. In cases where the sequenced isolate differs substantially from the available reference genome, mapping efficiency decreases due to sequence mismatches. This likely explains the low mapping rates observed for *T. indotineae* and *A. fumigatus* [22] . Additionally, non-specific primer binding can lead to off-target amplification, producing reads that do not correspond to the ITS region and therefore fail to align specifically [23]. Together, these factors demonstrate that high-quality mapping is not solely dependent on read quality, but also on the match between primers, the target region, and the available reference genomes.

Although high specificity may support accurate classification, the results showed that identification success did not consistently correlate with mapping percentage. For example, *C. albicans* had high specificity and abundance but was not correctly identified by Flye or wf-amplicon de novo. On the other hand, *T. indotineae* had low mapping and was not detected by GermGenie, yet was correctly identified by Wf-amplicon variant calling.

No consistent pattern was observed between the number of BLAST-detected primers, mapping specificity, relative abundance, and identification outcome. Some isolates with favorable input characteristics were misclassified, while others with lower-quality indicators were correctly identified. These findings suggest that differences in tool architecture, such as error correction and alignment strategy, have a stronger influence on identification success than read quality metrics alone.

This study has several limitations. The sample size was limited to ten isolates, and no actual FFPE-derived DNA was analyzed. Only one primer pair and a single fungal reference database were used. GermGenie was adapted for fungal ITS classification but was limited by database content and resolution. Other classifiers were not tested, and no evaluation was made of detection sensitivity in mixed samples.

Future studies could expand the dataset by including a broader range of fungal species, degraded DNA sources such as FFPE tissue, and alternative primer sets that reduce amplification bias. It would be useful to test GermGenie in combination with a custom fungal ITS database to improve classification, especially in cases where non-specific amplification produces off-target reads. For example, the UNITE database could be integrated to improve taxonomic coverage and classification accuracy. One way to address the challenge of genomic divergence is by including multiple reference genomes per species or strain group, rather than relying on a single representative. This would improve read mapping and variant calling in cases where the sequenced isolate differs from the available reference.

# CONCLUSION

This study showed that Nanopore sequencing of the fungal ITS region can be used for species-level identification of clinical isolates. Of all tested methods, the Wf-amplicon variant calling workflow in combination with ITSx and BLAST delivered the most accurate results. This method identified 9 out of 10 isolates correctly and proved to be the most reliable option. The tools Flye and wf-amplicon de novo were less consistent. They produced fewer usable consensus sequences and had lower identification success. This shows that not all reconstruction workflows are equally suitable for ITS-based identification. GermGenie was tested separately on the filtered read data. It allowed for quick screening but lacked precision for reliable species identification. The mapping results from Minimap2 and Samtools provided useful insight into PCR specificity. In several isolates, not all amplification was specific to the target region, as off-target reads were also detected. This shows that PCR performance may vary between fungal species. The outcomes of this project contribute to the development of fungal identification workflows that may also be applied to degraded DNA such as from FFPE samples. To improve accuracy further, better reference databases are needed.

# AVAILABILITY OF DATA AND MATERIALS

The datasets that were generated and analysed in this study are available. The full sequencing dataset, which consists of ten FASTQ files (one for each barcoded clinical fungal isolate), is available via Google Drive (<https://github.com/DaanAdriaanse/fungal-ITS-identification-pipeline/blob/main/DataAccess.md>).

All analysis scripts, workflow documentation, and example output files are publicly available in the GitHub repository (<https://github.com/DaanAdriaanse/fungal-ITS-identification-pipeline>).

# ACKNOWLEDGMENTS

The author would like to thank Dr. Eefje Schrauwen for her guidance and supervision throughout this research project. Additional thanks go to Sander Boden and Bazante Sanders for their valuable support and input during data analysis and interpretation.

# Reference

|  |  |
| --- | --- |
| [1] | W. Gu, . S. Miller en . C. Y. Chiu, „Clinical Metagenomic Next-Generation Sequencing for Pathogen Detection,” *National Library of Medicine,* 2019. |
| [2] | C. Muñoz-Cadavid , S. Rudd, S. R. Zaki, M. Patel en . S. A. Moser, „Improving Molecular Detection of Fungal DNA in Formalin-Fixed Paraffin-Embedded Tissues: Comparison of Five Tissue DNA Extraction Methods Using Panfungal PCR,” *National Library of Medicine,* 2010. |
| [3] | S. L. Mitchell en . P. J. Simner, „Next-Generation Sequencing in Clinical Microbiology: Are We There Yet?,” *Clinics in Laboratory Medicine,* vol. Volume 39, pp. 405-418, 2019. |
| [4] | L. F. Alviz, B. A. Jones en S. P. Agnihotri, „Identifying CNS infections in transplantation and immunomodulatory therapy,” vol. Aug 27, 2024. |
| [5] | C. L. Schoch, K. A. Seifert, . S. Huhndorf, V. Robert, . J. L. Spouge en W. Chen, „Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi,” *PNAS,* 2012. |
| [6] | E. Bellemain, T. Carlsen, . C. Brochmann, . E. Coissac, P. Taberlet en H. Kauserud, „ITS as an environmental DNA barcode for fungi: an in silico approach reveals potential PCR biases,” *BMC Microbiology,* 2010. |
| [7] | H. Do en A. Dobrovic, „Sequence artifacts in DNA from formalin-fixed tissues: causes and strategies for minimization,” *Clinical Chemistry,* vol. Volume 61, nr. Issue 1, p. 64–71, 2015. |
| [8] | M. Srinivasan, D. Sedmak en . S. Jewell, „Effect of Fixatives and Tissue Processing on the Content and Integrity of Nucleic Acids,” *The American Journal of Pathology,* pp. 1961-1971, 2002. |
| [9] | N. Quach, M. F. Goodman en D. Shibata, „In vitro mutation artifacts after formalin fixation and error prone translesion synthesis during PCR,” *BMC Clinical Pathology,* vol. Volume 4, 2004. |
| [10] | F. Pauly, R. Marienfeld, C. Miller en C. Schubert, „Damage induced by formalin fixation and paraffin embedding compromises genome integrity and reduces PCR amplification yield,” *Journal of Molecular Diagnostics,* pp. 295-303, 2011. |
| [11] | R. A. Samson en . J. Houbraken, „Isolation and identification of fungi.,” *Microorganisms in home and indoor work environments,* pp. 265-278, 2011. |
| [12] | R. N. Henrik, S. Anslan, M. Bahram, C. Wurzbacher, P. Baldrian en L. Tedersoo, „Mycobiome diversity: high-throughput sequencing and identification of fungi,” *Nature Reviews Microbiology,* p. pages 95–109, 2019. |
| [13] | „Nanoplot,” [Online]. Available: https://github.com/wdecoster/NanoPlot. |
| [14] | „Filtlong,” [Online]. Available: https://github.com/rrwick/Filtlong. |
| [15] | J. Bengtsson-Palme, Z. Wang, S. Branco en A. Godhe, „Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of,” *Methods in Ecology and Evolution,* 2013. |
| [16] | S. Boden, „GermGenie,” [Online]. Available: https://github.com/Avans-ATLS/GermGenie. |
| [17] | „Minimap2,” [Online]. Available: https://github.com/lh3/minimap2. |
| [18] | arq5x, „bedtools,” [Online]. Available: https://github.com/arq5x/bedtools/tree/master. |
| [19] | shokrofont, „Medaka,” [Online]. Available: https://github.com/nanoporetech/medaka. |
| [20] | R. Vaser , I. Sović en N. Nagarajan , „Fast and accurate de novo genome assembly from long uncorrected reads,” *Genome Research,* 2017. |
| [21] | K. Shafin, . T. Pesout , . R. Lorig-Roach, M. Haukness en C. Bosworth, „Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes,” *Nature Portfolio,* 2020. |
| [22] | K. Loit, K. Adamson, . M. Bahram , R. Puusepp en S. Anslan, „Relative Performance of MinION (Oxford Nanopore Technologies) versus Sequel (Pacific Biosciences) Third-Generation Sequencing Instruments in Identification of Agricultural and Forest Fungal Pathogens,” *Applied and Environmental Microbiology,* 2019. |
| [23] | E. Bellemain, T. Carlsen en C. Brochmann, „ITS as an environmental DNA barcode for fungi: An in silico approach reveals potential PCR biases. BMC Microbiology, 10, 189,” *BMC Microbiology,* 2010. |
| [24] | L. W. Hugerth en . A. F. Andersson, „Analysing Microbial Community Composition through Amplicon Sequencing: From Sampling to Hypothesis Testing,” *Frontiers,* vol. Volume 8 , 2017. |