Advancing Kitome Analysis in Low Abundance Sterility Testing: Blinded Microbial Study Through the Lens of Deep Learning and Nanopore Sequencing

James P. B. Strutt[1], M. Natarajan[1], Suiyuan Xia[1], Charles Swofford[2], Rohan B. H. Williams[1,3,4], Stacy L. Springs[1,2]

[1]Singapore-MIT Alliance for Research and Technology, Singapore.
[2]MIT Center for Biomedical Innovation, Massachusetts Institute of Technology, U.S.A.
[3]Singapore Centre for Environmental Life Sciences Engineering (SCELSE), Life Sciences Institute, National University of Singapore, Singapore
[4]SCELSE, Nanyang Technological University, Singapore.

1. Introduction

The detection of microbial contaminants in sterile cell therapy samples is critical for ensuring patient safety. However, kit-ome contamination—non-sample DNA originating from reagents, consumables, and analytical biases—complicates sequencing workflows by introducing false-positive signals. This issue is particularly challenging in low bio-load sterility testing, where accurate microbial detection is essential.

We present an approach that integrates sample preparation optimization, nanopore sequencing, deep learning, and standard metagenomics analysis to identify and mitigate kit-ome artifacts while enhancing microbial detection specificity. This pipeline improves sterility assessments by distinguishing true contaminants from background noise.

2. Methods and Implementation

2.1 Study Design and Data Collection

Blinded studies were conducted at a limit of detection (LOD) for USP-71 fungal and bacterial species spiked into human T-cells. rRNA and cpn60 were selected as universal microbial targets. Reads were analysed using metagenomic classification and sequence alignment, with nanopore sequencing serving as the endpoint detection method.

To establish baseline contamination profiles, negative control samples were analysed. The kit-ome baseline, alongside spiked organismal signal data, viral, and phage sequences, was then used to train and validate a deep learning pipeline designed to classify DNA sequences and separate true microbial contaminants from kit-ome artifacts.

2.2 Deep Learning Pipeline

Our approach employs the DNABERT-2-byte pair encoding (BPE) model [1], which extracts computational representations of DNA sequences for classification. The embeddings are processed by a custom transformer-based classifier optimized for detecting microbial and fungal signals amidst high background noise.

The model was trained on 8 million labeled reads and achieved the following performance on unseen data:

- Accuracy: 87.9% (±9.04%)
- Sensitivity: 0.855
- Precision: 0.896
- AUC: 0.988
- Loss: 0.36

This filtering step enabled the accurate identification of contaminant reads from amplicon sequences.

3. Related Work

Existing sterility testing methods rely on metagenomic sequencing and PCR-based assays, but struggle with background contamination from kit-ome artifacts [2]. Previous studies have explored long-read sequencing for microbial detection [3], yet deep learning applications for kit-ome artifact removal remain underdeveloped. Our approach combines sequencing, negative control profiling, and transformer-based deep learning to enhance detection specificity.

4. Results and Future Work

The proposed pipeline successfully distinguished contaminant organisms in spiked samples, reducing kit-ome sequence interference. Ongoing qPCR validation will confirm the method's robustness. Future efforts will focus on:

1. Scaling the method to additional sample and library preparation types.
2. Improving classification models for broader sterility testing applications.
3. Evaluating clinical and pharmaceutical applicability.

Acknowledgments

This research is supported by Agency for Science, Technology And Research (A*STAR) under the Industry Alignment Fund - Pre-positioning Programme (IAF-PP) grant: "Assembling Screening, Productions, High-throughput Analytics, and Lentiviral Targeting for T cells (ASPHALT) (H24J4a0031)" and National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme, through Singapore MIT Alliance for Research and Technology (SMART): Critical Analytics for Manufacturing Personalised-Medicine (CAMP) Inter-Disciplinary Research Group.

References

[1] Zhou *et al.,*. arXiv, 2024. DNABERT-S: Learning Species-Aware DNA Embedding with Genome Foundation Models.
[2] Salter *et al.,*. BMC Biology. 2014;12:87. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses.
[3] Strutt *et al.,*. Microbiology Spectrum (ASM). 2023;11(5). Machine-learning driven detection of adventitious agent detection in T cell cultures using long read sequencing.