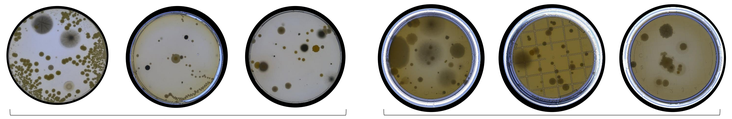
Application of Targeted Amplicon Next Generation Sequencing Using Curated Ribosomal Sequence Databases for Identification of Complex Microbial Samples

Lauren Salvitti, Daniel Kahline, Holly Anderson, Nick Bielski, Stella Loke, Prasanna Khot

1 ABSTRACT

Growth based methods for isolating microbes as pure subcultures, followed by characterization and/or identification (ID), is the standard practice in the industry for microbiome profiling applications, such as environmental monitoring and in-process testing. However, these methods can be time consuming if the subcultures contain mixed species, or are not compatible with current phenotypic, proteotypic, and genotypic ID methods resulting in a failure to generate actionable information. Targeted Amplicon Next Generation Sequencing (TA-NGS) provides an effective solution to identify microbes when current methods fail or are inadequate. TA-NGS amplifies either the bacterial 16S ribosomal gene or the fungal internal transcribed spacer gene – both well-established targets for microbial ID. By leveraging the high throughput nature of TA-NGS, thousands of sequence reads are produced per sample, allowing for resolution of mixed species samples when matched against a curated database for identification.



Settle plates

Contact plates

2 INTRODUCTION

There are several microbial ID methods available with varying levels of accuracy and resolution. These include growth-based methods based on phenotypic (colorimetric assays) and proteotypic targets (mass spectrometry). Genotypic methods, such as Sanger sequencing, provide the highest resolution without requiring additional growth, but depend on isolation to provide a definitive ID. These approaches successfully provide IDs for most samples; however, reliance on growth and/or isolation extends result timelines and requires organisms to be culturable. In addition, species with complex poly insertion/deletion events in their multicopy ribosomal genes result in differences of target sequence length, often making sequence assembly difficult and/or unreliable. In contrast to Sanger sequencing, Next Generation Sequencing (NGS) is a massively parallel sequencing technology able to generate millions of reads in a single run. Resulting reads are demultiplexed and matched against a database to provide IDs of all organisms present in a sample. Thus, NGS technologies generate sequencing reads from mixed species samples without any need for additional time-consuming growth or isolation steps.

A bioinformatic pipeline integrated with a highly curated ribosomal gene sequence database for the TA-NGS workflow to identify bacterial and fungal samples was developed. The performance of this pipeline was compared to Sanger sequencing-based IDs.

3 METHODS

A dataset was generated from 147 samples (77 bacterial and 70 fungal) consisting of single species and simulated mixes of 2-4 species with identifications verified by Sanger Sequencing (Figure 1). This dataset was used to optimize TA-NGS data interpretation thresholds for reporting results. An additional 18 samples representing complex poly-insertion/deletion events, which were unable to obtain an identification via Sanger sequencing, were included for testing (Figure 1). Genomic DNA was first extracted from each sample before being amplified using primers targeting the 16S gene (bacteria) and ITS2 gene (fungi/yeast) (Figure 2).

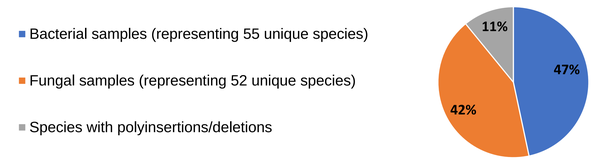


Figure 1. Breakdown of samples included in cohort.

METHODS (cont.)

Barcode sequences were then ligated to individual samples, pooled, and finally sequenced with 2x300 chemistry (Figure 2). Raw reads were trimmed for quality, paired, and filtered for chimeric sequences followed by analysis via an Amplicon Sequence Variant (ASV) pipeline in conjunction with the curated 16S and ITS libraries. To assess the performance of the pipeline in providing identifications, we evaluated the accuracy of the classifications compared to Sanger sequence-based IDs, and sensitivity relative to the number of species present in each sample.

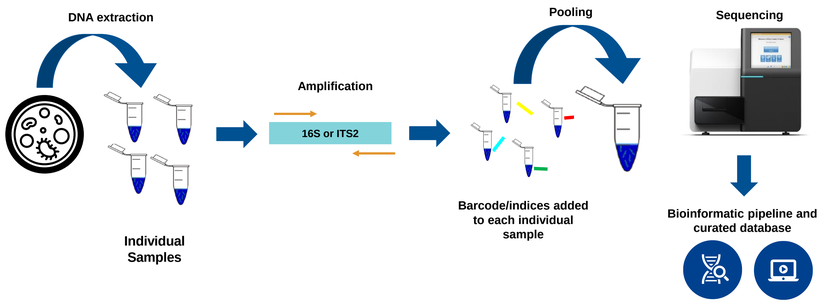


Figure 2. Targeted Amplicon Sequencing of the 16S or ITS2 region.

4 RESULTS

A total number of 251 classification calls ranging from species to kingdom taxonomic levels were generated from 147 bacterial and fungal samples. For each classification call made at the genus or species level, a comparison was made with the Sanger based reference ID to determine if the call was a match or a non-match. Data interpretation thresholds were set for the sequence read output of the bioinformatic pipeline to maximize ID rate and accuracy (Table 1). For the 18 samples that were unable to be sequenced via Sanger sequencing, there was a 100% ID rate via TA-NGS with the selected thresholds.

- Accuracy (overall correctness of calls) = True Positives / (True Positives + False Positives)
- Sensitivity (ability to detect all species in a sample) = True positives / (True positives + False negatives)

Table 1. Output metrics for bacterial and fungal datasets. Accuracy and sensitivity are reported for species and genus level calls only.

Dataset	ID rate	Accuracy*	Sensitivity
Bacterial (n=77)	100%	95.56%	98.38%
Fungal (n=70)	100%	98.21%	96.07%

*False positives are attributed to species level mismatches of closely related species within the same genus.

Table 2. Example of thresholded output for a mixed bacterial sample (n=2) and a sample with a poly insertion/deletion.

Sample	Reference ID	TA-NGS Output Taxon	Rank	% Seq. Reads
Mixed species (n=2)	*Corynebacterium aurimucosum*	*Corynebacterium aurimucosum*	species	52.76%
	Kocuria indica / marina	*Kocuria*	genus	47.24%
Species with poly insertion/del.	No ID	*Corynebacterium tuberculostearicum*	species	86.27%
		Corynebacterium	genus	12.96%
		Classifications below threshold	N/A	0.77%

CONCLUSIONS

Accurate and fast microbial identifications are a cornerstone of a successful EM program and in-process testing. TA-NGS serves as a solution for generating microbial IDs for samples that fail with current methods due to mixed species or complex poly-insertion/deletion events. In addition, TA-NGS can be an option for sequencing mixed species samples without the need for further isolation.

A TA-NGS bioinformatic pipeline was developed to identify bacterial and fungal single and mixed species samples:

- A threshold was optimized for data interpretation to maximize accuracy and sensitivity.
- The pipeline had an ID rate of 100%, and average accuracy of 96.9% and sensitivity of 97.2% for both bacterial and fungal samples. This was achievable due to integration of the pipeline with a highly curated Sanger sequencing library.
- The small percentage of false positive calls were attributed to species level mismatches of closely related organisms within the same genus.
- Samples with poly in/del were identified at a 100% ID rate. The majority of these samples were part of the *Corynebacterium* or *Methylobacterium* genus.