16S SNAP APP: An automated pipeline for community analysis using multiple 16S rRNA variable regions

INTRODUCTION

Microbial profiling through sequencing of the 16S rRNA gene region has greatly advanced our understanding of microbial evolution and the relationship between microbial diversity and its important roles in human and environmental health. Traditional workflows use sequences generated from only one or two of the nine 16S rRNA variable regions to determine taxon abundances and classifications. By ignoring the remaining variable regions, a substantial amount of compositional data is lost. The **xGen™ 16S Amplicon Panel v2** (Catalog No. 10009828, formerly known as Swift Normalase™ 16S v2 Panel) solves this problem by providing efficient and comprehensive profiling of all nine 16S rRNA variable (V) regions in a single reaction (**Figure 1**).

To leverage full potential of the information from multiple V-regions captured by the xGen 16S Amplicon Panel v2 libraries, we designed the 16S SNAP APP (16S SNAPP), which incorporates additional steps to associate sequence reads from different V-region amplicons to generate consensus sequences for taxonomic classification. In this application note, we describe 16S SNAPP, a user-friendly, multiple V-regions aware, automated pipeline for generating high resolution microbial community profiles from xGen 16S Amplicon Panel v2 library sequencing data. This open-source analysis tool is available at https://github.com/swiftbiosciences/16S-SNAPP-py3.

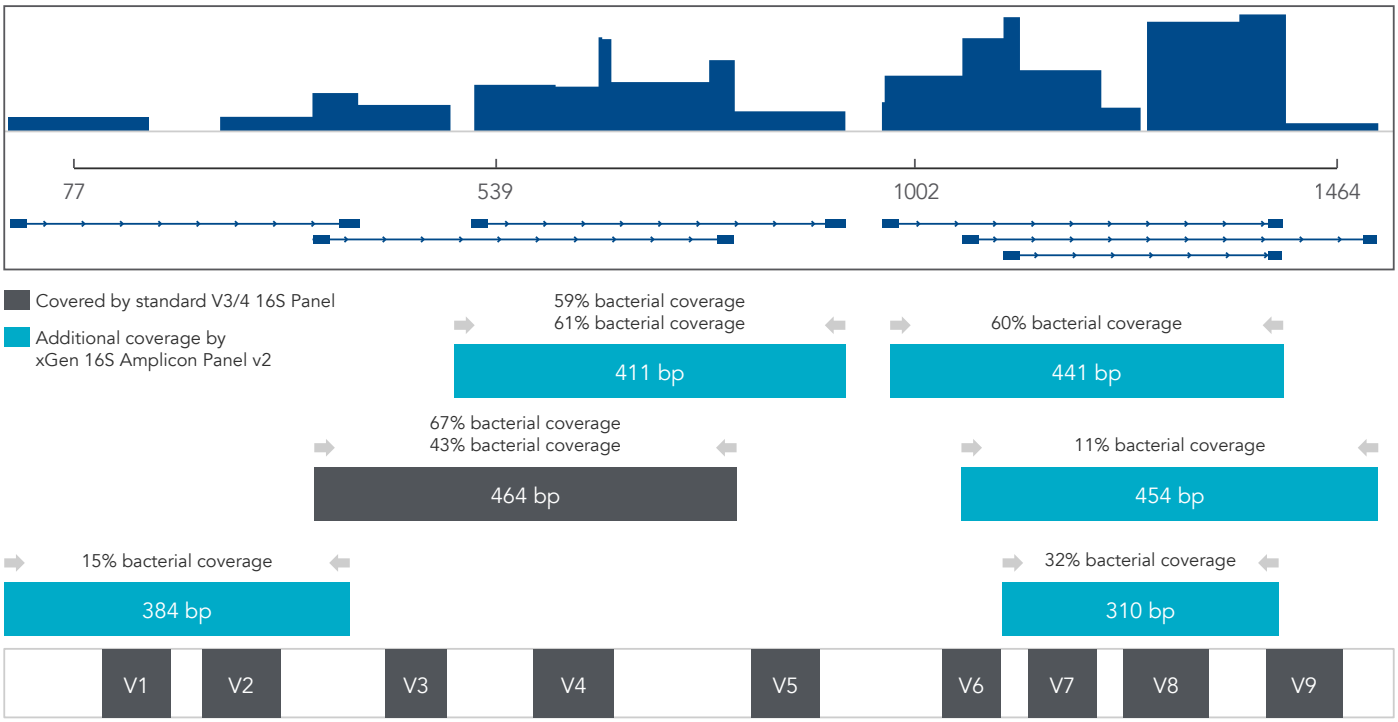

Figure 1. xGen 16S Amplicon Panel v2 (primers). Sequencing read coverage for an *E. coli* DNA sample (n=1) observed in Integrative Genomics Viewer (IGV) Sashimi plot and illustration of multiplexed amplicon coverage of all nine variable regions of 16S rRNA (gray and light blue bars) compared to a standard V3/V4 16S sequencing read coverage (gray bar only).

PIPELINE OVERVIEW

While data produced from xGen 16S Amplicon Panel v2 libraries can be processed using traditional 16S rRNA sequencing software, including mothur [1], QIIME 2™ [2], and RDP [3], these tools were designed for analysis of a read pair covering the entirety of only one or two contiguous variable regions. Without any mechanism to associate amplicon reads from all nine variable regions, these tools can only perform read-level classification which can potentially obscure community composition. For example, a bacterial taxon present in the sample may not be detected if none of its targeted regions alone can be assigned to that taxon in absence of additional information from other regions.

16S SNAPP workflow is designed for paired-end data of commonly used Illumina® sequencing lengths. The examples here show high-resolution classification from PE150 sequenced reads though PE300 can be readily processed as well. The pipeline starts with primer trimming using Cutadapt [4] followed by quality filtering, denoising, paired-end merging (using 'justConcatenate' option since not all read pairs have overlaps), and chimera removal using DADA2 [5] resulting in amplicon sequence variants (ASVs) (Figure 2). ASV read pairs are then split into single reads and undergo dereplication using VSEARCH [6] to form a smaller, unique sequence set, which is used to query the custom 16S database (RDP 11.5, https://rdp.cme.msu.edu) using BLAST [7] for high identity matches. By curating matching target sequences to maximize alignment coverages and read counts, 16S SNAPP enriches multiple variable regions to a minimum set of reference sequences (the template set). This collection of sequences is assumed to encompass the template sequences amplified during library generation. From template-aligned read pairs consensus sequences are computed. These sequences, together with sequence features of individual non-aligned read pairs, are classified using RDP Classifier 2.13 [8] to generate and output a standard format abundance table. Template sequences are used to construct reference trees using MAFFT [9] and FASTTREE [10]. With the ability to associate sequences from different amplicon regions to form consensus sequences, 16S SNAPP generates improved classification beyond read-level analyses.

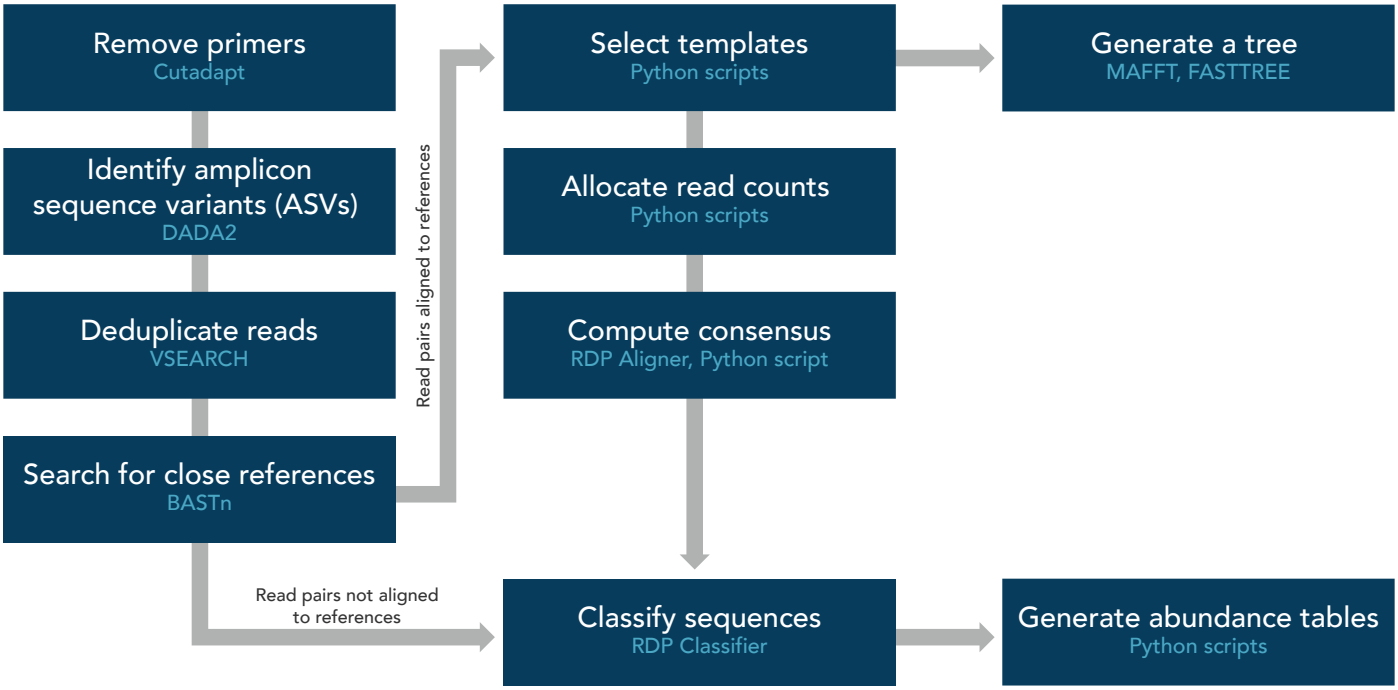

Figure 2. Diagram depicting the 16S multi-amplicon analysis workflow for data generated using the xGen 16S Amplicon Panel v2.

PERFORMANCE METRICS

Output of the 16S SNAPP workflow was compared to read-level analysis implemented with RDP and QIIME 2 to demonstrate the added value of using multiple variable regions relative to the traditional 16S rRNA sequencing analysis methods. Performance was assessed using the following parameters: assignment accuracy, classification resolution, and run time performance.

Assignment accuracy

Compared to RDP and QIIME 2, 16S SNAPP assigned a substantially higher percentage of reads to expected genera from two mock bacterial communities—MSA-1003™ (ATCC®, 20 bacterial species) and ZymoBIOMICS™ Microbial Community Standard I & II (Zymo Research Catalog Nos. D6305, D6311; 8 bacterial species) (Figure 3A and 3B). Additionally, observed genome abundances were more highly correlated with expected abundances compared to RDP and QIIME 2 (Figure 4A and 4B). These results indicate increased true positive assignment rates and more accurate abundance estimates using the 16S SNAPP workflow.

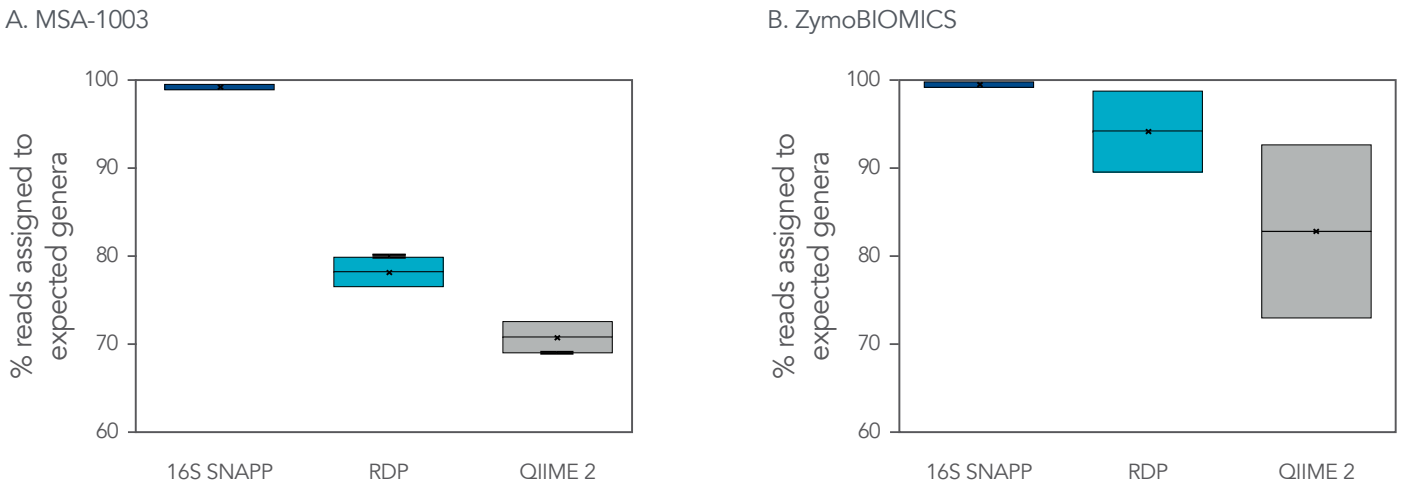

Figure 3. 16S SNAPP provides superior genus-level assignment of commercially available microbial community standards. xGen 16S Amplicon Panel v2 libraries were prepared using genomic DNA standards from mock bacterial communities, (**A**) MSA-1003™ (ATCC®) and (**B**) ZymoBIOMICS™ Microbial Community DNA Standard I & II (Zymo Research®). Sequencing data was analyzed using 16S SNAPP, RDP, or QIIME 2™. Since bacterial composition is known in the standards, percent of reads assigned to expected genera were determined for each pipeline. Standard deviations were calculated from four replicate libraries of each standard.

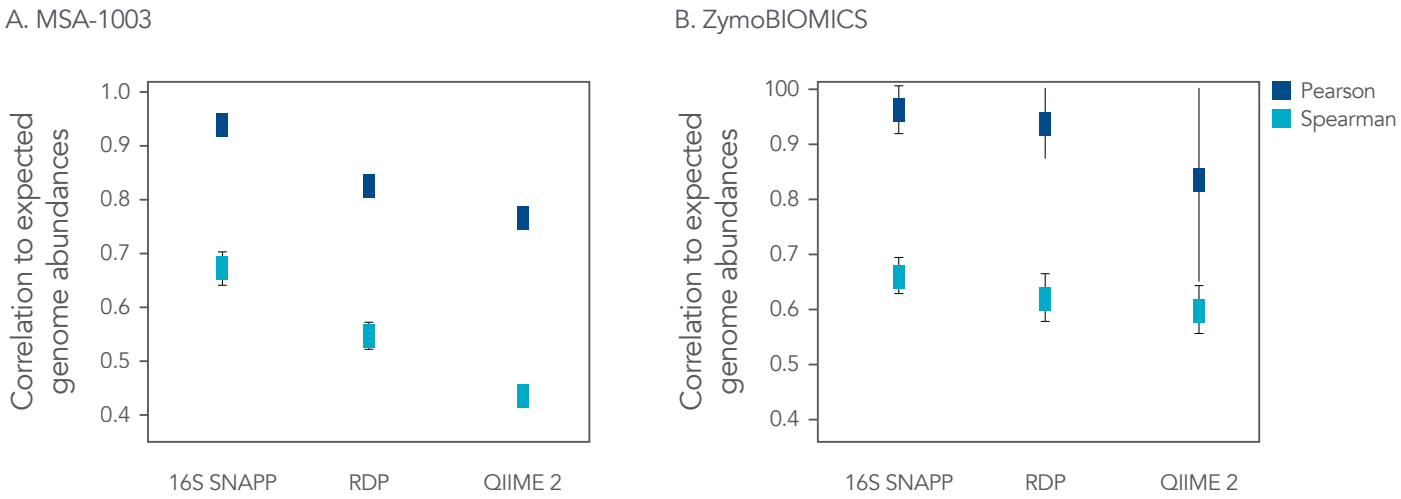

Figure 4. Correlation of observed to expected microbial abundances determined by 16S SNAPP, RDP, and QIIME 2™. xGen 16S Amplicon Panel v2 libraries were prepared using genomic DNA standards from mock bacterial communities, (**A**) MSA-1003™ (ATCC®) and (**B**) ZymoBIOMICS™ Microbial Community DNA Standard I & II (Zymo Research®). Sequencing data was analyzed using 16S SNAPP, RDP, or QIIME 2. Pearson and Spearman correlation coefficients between observed and expected genome abundances were determined for each pipeline.

To further analyze 16S SNAPP, we generated data from high-complexity, swine manure samples prepared with the xGen 16S Amplicon Panel v2 panel. Results show that 16S SNAPP outperformed both RDP- and QIIME-based workflows by assigning higher percentages of reads to genus-level taxa (Figure 5). The increased taxonomic resolution suggests that 16S SNAPP characterizes community composition to a greater degree than traditional read-level only analyses.

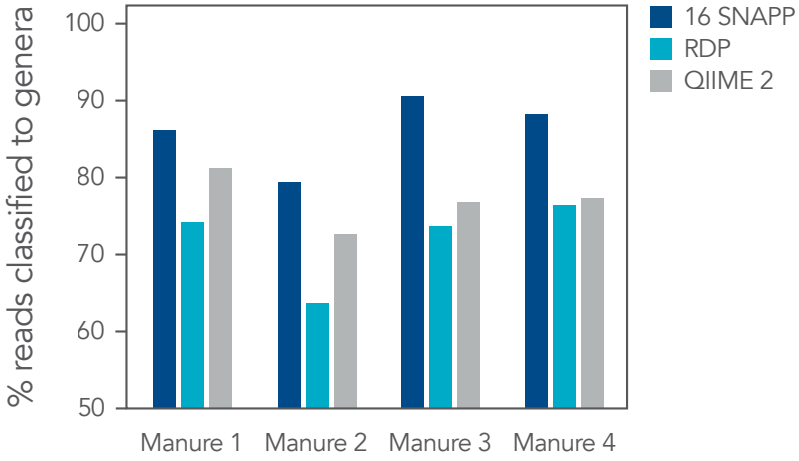

Figure 5. Genus-level assignment performance of complex bacterial community samples. xGen 16S Amplicon Panel v2 libraries were prepared using genomic DNA from high-complexity swine manure samples (n=4). Sequencing data was analyzed using 16S SNAPP, RDP, or QIIME 2™. Percent of reads assigned to 157 bacterial genera were determined for each pipeline.

Run time performance

For compute time evaluation of 16S SNAPP, we found that the pipeline scales linearly with increasing sample complexity, demonstrating consistent performance. For example, a ZymoBIOMICS Microbial Standard I sample of 89K reads (8 bacterial species, 362 unique ASVs) takes 10 minutes to complete, while a swine manure sample of 1.3 million reads (7474 unique ASVs) requires 162 minutes (Figure 6).

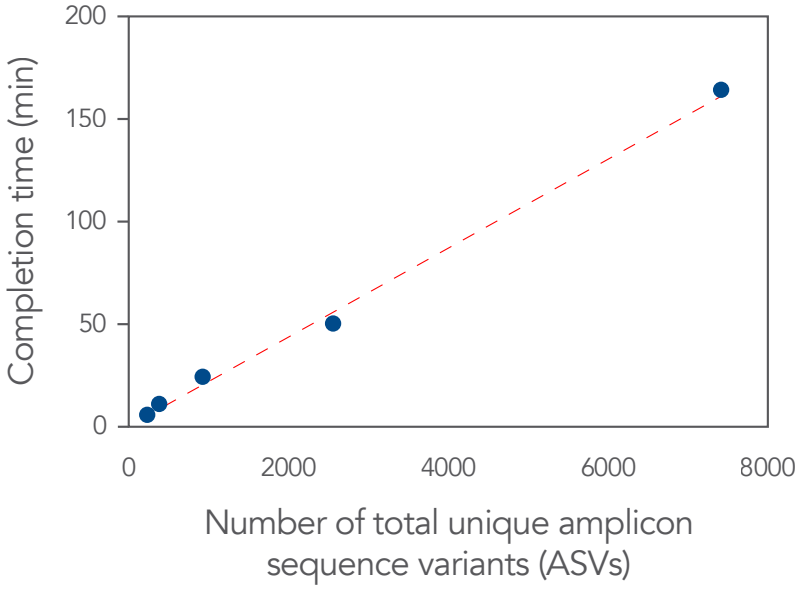

Figure 6. 16S SNAPP computing run time performance directly correlates with sample complexity. Computational completion time through 16S SNAPP pipeline was determined for one sample each from five different data sets (n=5) of varying community complexity, as measured by number of unique amplicon sequence variants.

REFERENCES

1. Schloss PD, Westcott SL, Ryabin T, et al. **Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities.** *Appl Environ Microbiol.* 2009;75(23):7537-7541. doi:10.1128/AEM.01541-09
2. Bolyen E, Rideout JR, Dillon MR, et al. **Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2.** *Nat Biotechnol 2019 378.* 2019;37(8):852-857. doi:10.1038/s41587-019-0209-9
3. Cole JR, Wang Q, Fish JA, et al. **Ribosomal Database Project: data and tools for high throughput rRNA analysis.** *Nucleic Acids Res.* 2014;42(D1):D633-D642. doi:10.1093/NAR/GKT1244
4. Martin M. **Cutadapt removes adapter sequences from high-throughput sequencing reads.** *EMBnet.journal.* 2011;17(1):10-12. doi:10.14806/EJ.17.1.200
5. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. **DADA2: High-resolution sample inference from Illumina amplicon data.** *Nat Methods 2016 137.* 2016;13(7):581-583. doi:10.1038/nmeth.3869
6. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. **VSEARCH: a versatile open source tool for metagenomics.** *PeerJ.* 2016;4(10). doi:10.7717/PEERJ.2584
7. Camacho C, Coulouris G, Avagyan V, et al. **BLAST+: architecture and applications.** *BMC Bioinformatics.* 2009;10. doi:10.1186/1471-2105-10-421
8. Wang Q, Garrity GM, Tiedje JM, Cole JR. **Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy.** *Appl Environ Microbiol.* 2007;73(16):5261-5267. doi:10.1128/AEM.00062-07
9. Katoh K, Misawa K, Kuma KI, Miyata T. **MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.** *Nucleic Acids Res.* 2002;30(14):3059-3066. doi:10.1093/NAR/GKF436
10. Price MN, Dehal PS, Arkin AP. **FastTree: computing large minimum evolution trees with profiles instead of a distance matrix.** *Mol Biol Evol.* 2009;26(7):1641-1650. doi:10.1093/MOLBEV/MSP077

APPENDIX

Hardware and software requirements

The below requirements provide the minimum computing resources required to run 16S SNAPP. For ease of use, the workflow and all dependencies are provided on GitHub.

- Linux or Mac command line interface
- 32 GB of RAM minimum
  Note: Memory usage is highly dependent on sample number and complexity; specific applications may require more
- Multi-core processors are recommended to enable parallel processing
- R ≥3.6.0 (https://www.r-project.org/)
- Java ≥1.8.0_131
- Python 3.6.8 with Numpy ≥1.16.2, Pandas ≥1.1.14, and Scipy ≥1.4.1
- DADA2 (https://benjjneb.github.io/dada2/)
- BLAST+ (https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/)
- RDPTools (https://github.com/rdpstaff/RDPTools)
- Cutadapt (https://cutadapt.readthedocs.io/en/stable/)
- VSEARCH (https://github.com/torognes/vsearch)
- MAFFT (https://mafft.cbrc.jp/alignment/software/)
- FASTTREE (http://www.microbesonline.org/fasttree/)
- minimap2 ≥2.20-r1061 (https://github.com/lh3/minimap2)

16S SNAP APP: An automated pipeline for community analysis using multiple 16S rRNA variable regions

Technical support: applicationsupport@idtdna.com

Doc ID: RUO22-0711_001 04/22