Critically Assessing the Influence of Variable Region Selection on Bacterial Community Profiling: 16S rDNA and the sequencing.

Rust Turakulov[1,2], Lesley Gray[1], Jake Robinson[2], Martin Breed[2], Rebecca Jordan[3], Naga Kasinadhuni[1], Melanie O'Keefe[1], Jack Royle[1], Nathan Bachmann[1], Martha Zakrzewski[1], Trent Peters[1], Kelsey Maloney[1], Shan Zong[1], Christopher Noune[1], Matthew Tinning[1], John Stephen[1], Cath Moore[1].

[1] AGRF: Victorian Comprehensive Cancer Centre, 305 Grattan St, Melbourne, VIC 3000

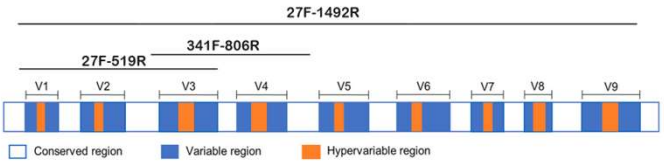
[2] Flinders University: College of Science & Engineering, Sturt Road, Bedford Park South Australia 5042

[3] CSIRO Environment: 15 College Rd, Sandy Bay, Tasmania 7005 Contact: Rust.Turakulov@agrf.org.au

Introduction: variable regions and samples

The 16S rDNA gene is highly conserved across different species of bacteria but contains hypervariable regions that provide species-specific signature sequences useful for bacterial identification. The 16S rRNA amplicon sequencing is widely used for detecting and characterizing microbial species in microbial communities. Taxonomic classification can be influenced by the choice of the 16S variable region. In this study, we compared the outcomes of three sequencing experiments on 30 soil samples from a mining revegetation site, using three different regions:

- V1-V3 with the Illumina: 27F-519R 492bp
- V3-V4 with the Illumina: 341F-806R 465bp
- Full-length 16S with the PacBio: 27F-1492R 1465bp

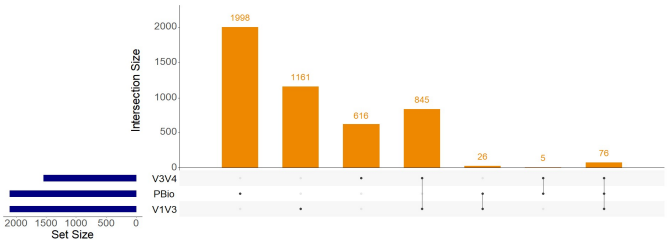


Number of detected bacterial species

Despite the common expectation that the longer target should produce better species resolution, we found V1-V3 region resolved a similar number of taxa to the full-length sequence.

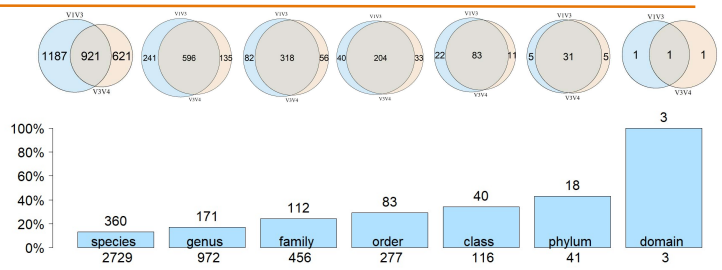
- V1-V3 Set Size: n = 2,108 (1,161 unique: ~55%)
- V3-V4 Set Size: n = 1,542 (616 unique: ~40%)
- PacBio Set Size: n = 2,105 (1,998 unique: ~95%)

However, the best overlap in species detection was observed for Illumina platform (V1-V3 vs V3-V4) while very little of taxonomy profiles shared between PacBio and Illumina experiments: only 76 of the same species were called across all three experiments. One of the likely reasons which potentially lower efficiency of PacBio workflow to identify species level is fact that over 40% reads were put to unclassified taxonomy bin due to the absence of full-length reference data in the available databases.



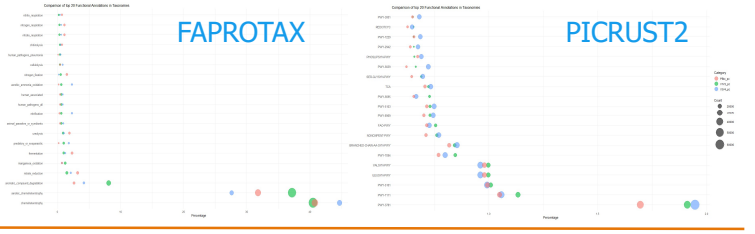
Differentially abundant taxonomies

To compare differences in taxonomic assignment between the Illumina protocols using V1-V3 versus V3-V4 amplicons, we analyzed the differential abundance of bacterial profiles at seven taxonomic levels using a paired t-test with FDR adjustment and a p-value threshold of 0.05. Interestingly, we found that the higher the taxonomic level cutoff, the greater the proportion of differentially abundant bacterial entities between the V1-V3 and V3-V4 protocols. In other words, the closest match in profiles was observed at the species taxonomic level, while matches in abundance decreased at higher taxonomic levels. For example, at the domain level, there are only three groups: 1—Bacteria, 2—Archaea, and 3—Unclassified. Archaea were only observed with the V3-V4 protocol, and unclassified entities were present in the V1-V3 protocol, indicating that the V1-V3 region lacks sufficient variability to distinguish the Archaea group. Intermediate taxonomic levels showed more overlap (as indicated by the Venn diagrams above the bars); however, we observed a steady increase in the number of entities with significant differences in abundance identified by the different amplicons.



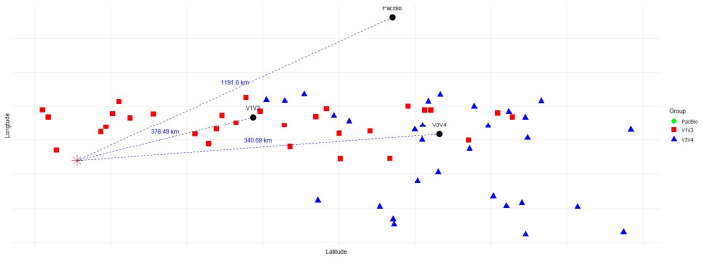
Functional analysis of bacterial communities

Despite substantial differences in bacterial species identification, the functional (metabolic activity) predictions of bacterial communities, performed using the FAPROTAX and PICRUSt workflows, demonstrated a similar pattern across all three amplicons. The three most abundant pathways were the same for all amplicons. The top pathway identified by the FAPROTAX program was "chemoheterotrophy," which derives energy from the oxidation of organic molecules, such as glucose, proteins, and lipids, with the primary source of carbon coming from pre-existing organic compounds. This functional class accounted for over 40% of the taxonomic hits. In contrast, with PICRUSt, the pathways database is more granular, and the main pathway, "PWY-3781," accounted for 1-2% of the bacterial community. This pathway, defined on *BioCyc.org* as aerobic respiration I (cytochrome c), is similar to mitochondrial respiration in mammals and plants.



Geographical location prediction

The Australian Microbiome Initiative (AMI) provides a publicly available 16S rDNA dataset of over 10,000 samples collected across Australia, accompanied by rich metadata. This resource can be utilized to train machine learning models to predict metadata based on bacterial communities. The AMI dataset includes V1-V3 amplicon sequences for over 3,000 soil samples, which we used to train a randomForest regression model to predict latitude and longitude for unknown soil samples. As expected, the V1-V3 amplicon data provided the best accuracy for prediction, with an average error of 376.40 km from the true site, indicated by the star on the figure on the right relative to the V1V3 group center. The center of the V3-V4 group has a similar average error of 340.68 km from the actual collection site; however, this blue triangle group shows much greater dispersion in longitude compared to the red square V1V3 dataset, which demonstrated more compact grouping. The prediction model using the PacBio full-length sequences returned a single coordinate located 1191.6 km from the true site. This discrepancy likely reflects the minimal overlap between the species used for model training with the PacBio pipeline and those in the SilvaV138 database used for the Illumina dataset.



Conclusions

When selecting an amplicon for 16S rDNA bacterial profiling, consider the following critical factors based on the project's objectives:

1) **V1-V3:** This protocol identifies the most species, with fewer than 5% of reads remaining unclassified. It is a stable protocol that has been refined over time and has accumulated a substantial amount of historical data, making it potentially suitable for a wide range of experiments. The optimized protocol provides consistent results with minimal sample-to-sample variability. However, this amplicon cannot distinguish Archaea.

2) **V3-V4:** This protocol offers even better performance and consistency in sequencing metrics due to shorter amplicon sizes. It can detect some Archaea species and can be integrated with V1-V3 datasets with minimal bias. However, because of the shorter amplicon, less sequence variability is captured, leading to fewer species-level taxonomies being accurately distinguished. The public dataset for V3-V4 is not as extensive as for V1-V3 amplicons. moment there is limited historical data accumulated for this protocol.

3) **PacBio:** Economically, PacBio offers an advantage over Illumina amplicon sequencing, especially for relatively small projects. However, a significant drawback is the current long-read pipeline, which fails to identify up to 40% of reads from soil bacteria. That effect is related to the reference sequences database which lack full length amplicons. The remaining reads yield a similar number of unique taxonomies to V1-V3 amplicon, although the number of ASVs (Amplicon Sequence Variants) is much higher than what would be detected with Illumina amplicons. The PacBio protocol is relatively new and subject to further evolution, which may affect future data integration. Additionally, at this moment there is limited historical data accumulated for this protocol.

4) **Functional Analysis:** Functional analysis using PICRUSt2 and FAPROTAX shows remarkable reproducibility across all types of amplicons, indicating no significant advantage of one protocol over another.

5) **Prediction Modelling:** Prediction modeling performs best when the same type of amplicon is used for both the model training data and the project dataset. When planning a project with machine learning algorithms in mind, it is crucial to select a stable dataset and a matching protocol that can be reused in the future. For soil data, the V1-V3 protocol is most recommended.

Acknowledgements

This project was supported by the Cooperative Research Centre for Transformations in Mining Economies.

www.agrf.org.au