Profiling Metagenomic Communities Using Circular Consensus and Single Molecule, Real-Time Sequencing

Richard J. Hall[1], Cheryl A. Gale[2], Matt Hamilton[3], Tim Heiseland[3], Alex Knotted[3], Michael Sadowsky[3], Kevin Silverstein[4], Alexa Weingarden[3] and Cheryl Heiner[1]

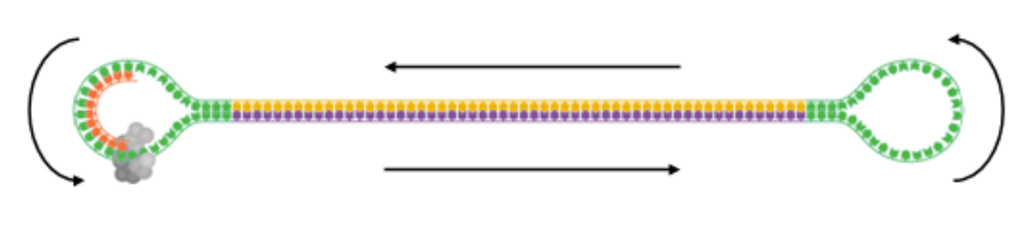
[1]Pacific Biosciences, Menlo Park, CA, [2]Department of Pediatrics, University of Minnesota, [3]BioTechnology Institute, University of Minnesota, [4]Supercomputing Institute for Advanced Computational Research, University of Minnesota

Introduction

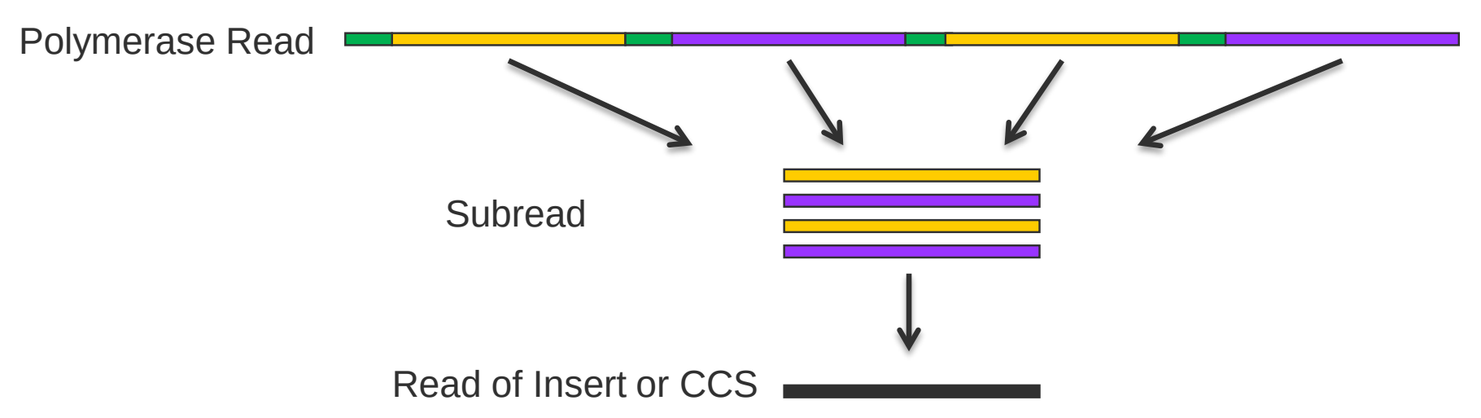
There are many sequencing-based approaches to understanding complex metagenomic communities spanning targeted amplification to whole-sample shotgun sequencing. While targeted approaches provide valuable data at low sequencing depth, they are limited by primer design and PCR amplification. Whole-sample shotgun experiments generally use short-read, second-generation sequencing, which results in data processing difficulties. For example, reads less than 1 Kb in length will likely not cover a gene or region of interest, and will require assembly. This not only introduces the possibility of incorrectly combining sequence from different community members, it requires a high depth of coverage. As such, rare community members may not be represented in the resulting assembly.

Circular-consensus, single molecule, real-time (SMRT®) Sequencing reads in the 1-2 kb range, with >99% accuracy can be efficiently generated for low amounts of input DNA. 10 ng of input DNA sequenced in 4 SMRT Cells would generate >100,000 such reads. While throughput is low compared to second-generation sequencing, the reads are a true random sampling of the underlying community, since SMRT Sequencing has been shown to have no sequence-context bias. Long read lengths mean that that it would be reasonable to expect a high number of the reads to include gene fragments useful for analysis.

Profiling Using Circular Consensus Reads



SMRTbell™ libraries were prepped from whole-sample metagenomes that were each sheared to target size of 2 Kb. Libraries were sequenced using P6/C4 chemistry with 3-hr collection times.

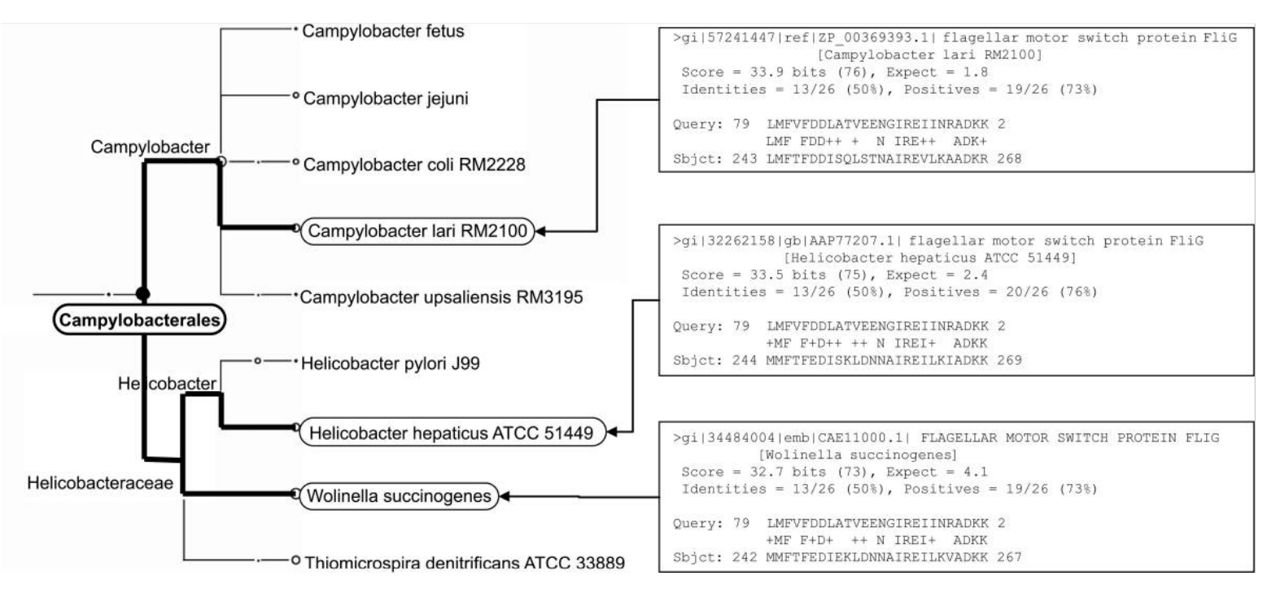


Multiple sequencing passes were generated from each SMRTbell template, generating high quality circular consensus sequence (CCS) reads.

CCS read with predict genes
Translate
Protein sequences

Prodigal (**Pro**karyotic **D**ynamic **P**rogramming **G**enefinding **Al**gorithm)[1], is used to predict genes from individual CCS reads and impute amino acid sequences.

blastp is used to align the putative protein sequences to the RefSeq (http://www.ncbi.nlm.nih.gov/refseq/) bacterial database


Campylobacter fetus
Campylobacter jejuni
Campylobacter
Campylobacter coli RM2228
Campylobacter lari RM2100
Campylobacter upsaliensis RM3195
Campylobacterales
Helicobacter pylori J99
Helicobacter
Helicobacter hepaticus ATCC 51449
Helicobacteraceae
Wolinella succinogenes
Thiomicrospira denitrificans ATCC 33889

```
>gi|57241447|ref|ZP_00369393.1| flagellar motor switch protein FliG
                [Campylobacter lari RM2100]
 Score = 33.9 bits (76), Expect = 1.8
 Identities = 13/26 (50%), Positives = 19/26 (73%)

Query: 79  LMFVFDDLATVEENGIREIINRADKK 2
            LMF FDD++ +  N IRE+  ADK+
Sbjct: 243 LMFTFDDISQLSTNAIREVLKAADKK 268
```

```
>gi|32262158|gb|AAP77207.1| flagellar motor switch protein FliG
                [Helicobacter hepaticus ATCC 51449]
 Score = 33.5 bits (75), Expect = 2.4
 Identities = 13/26 (50%), Positives = 20/26 (76%)

Query: 79  LMFVFDDLATVEENGIREIINRADKK 2
           +MF F+D++ ++ N IREI+  ADK
Sbjct: 244 MMFTFEDISKLDNNAIREILKIADKK 269
```

```
>gi|34484004|emb|CAE11000.1| FLAGELLAR MOTOR SWITCH PROTEIN FLIG
                [Wolinella succinogenes]
 Score = 32.7 bits (73), Expect = 4.1
 Identities = 13/26 (50%), Positives = 19/26 (73%)

Query: 79  LMFVFDDLATVEENGIREIINRADKK 2
           +MF F+D+  ++ N IREI+  ADK
Sbjct: 242 MMFTFEDIEKLDNNAIREILKVADKK 267
```

Blast results are imported into MEGAN[2] and a Lowest Common Ancestor (LCA) algorithm is used to assign a taxonomy to each putative protein sequence.

BEI Mock Staggered Community

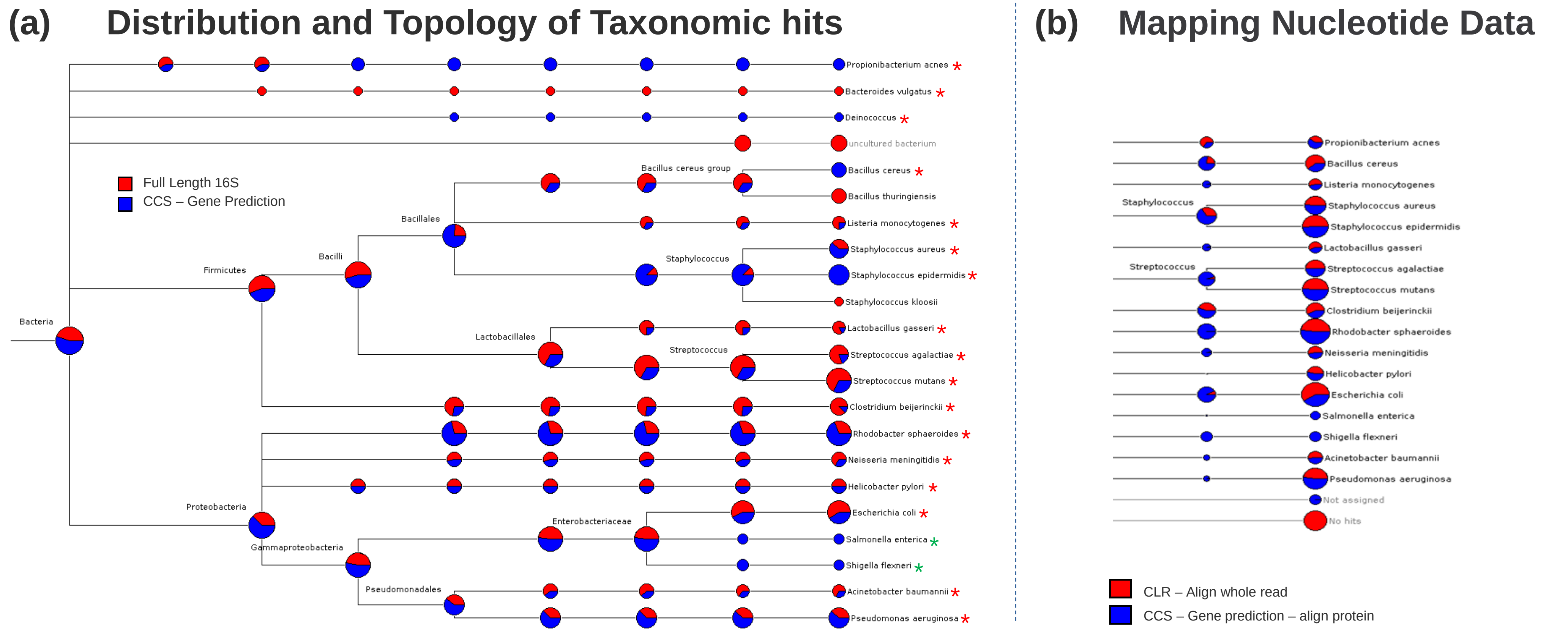


Figure 1. Comparison of CCS-gene prediction and other mapping strategies (a) Distribution of taxonomic hits to the Silva Database[4] using CCS gene predictions (blue) and 16S sequences (red). Filtered at 0.0005% abundance. Pie charts on nodes show total of all downstream assignments. * Indicates known members of the mock community. * Incorrect assignment likely due to gene sharing between *Salmonella*, *Shigella* and *E. coli*. **(b)** Mapping predicted genes gives comparable abundance numbers to mapping of raw sequence reads.

Stool Sample from a 4 Day Old Healthy Infant

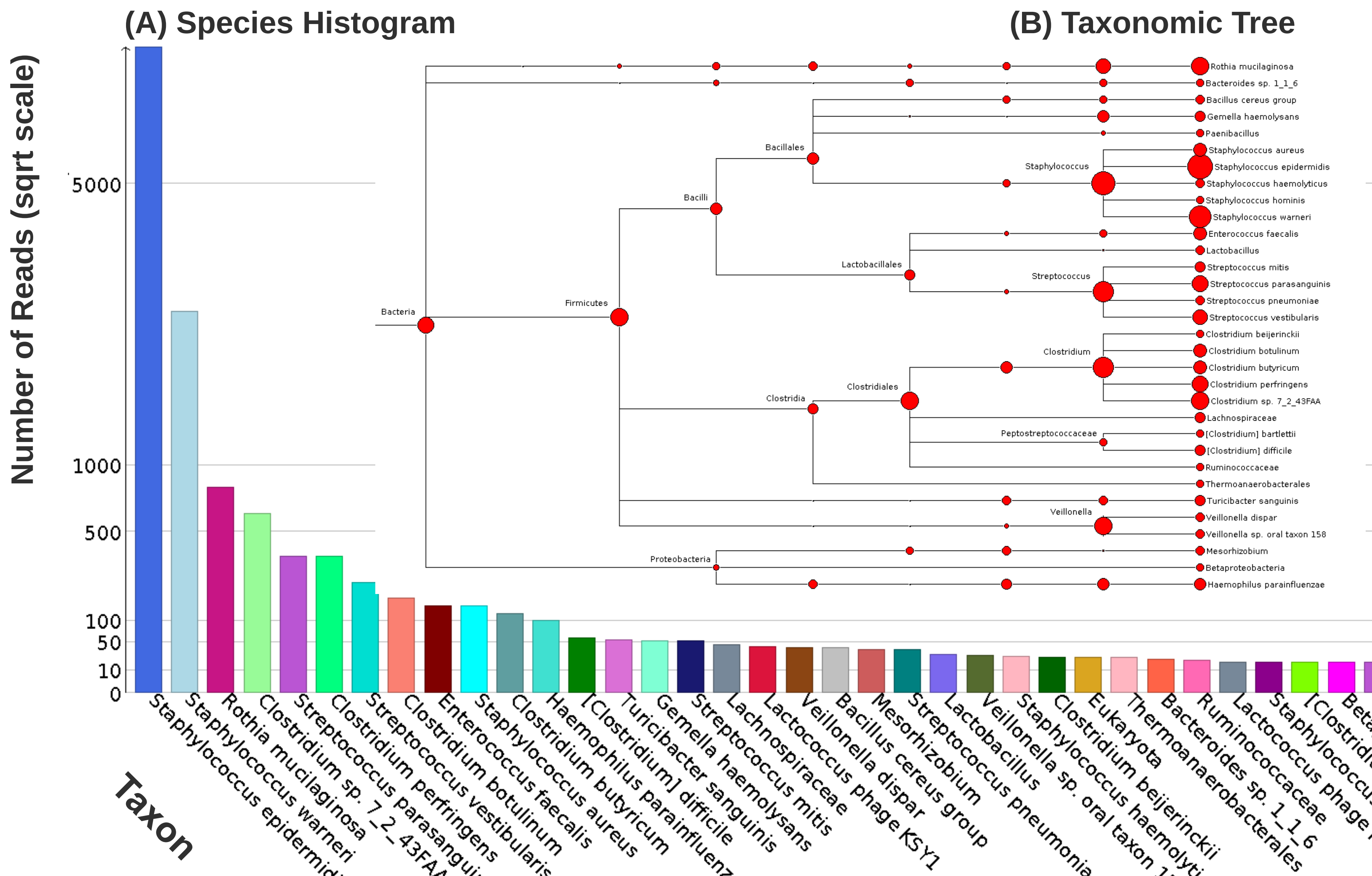
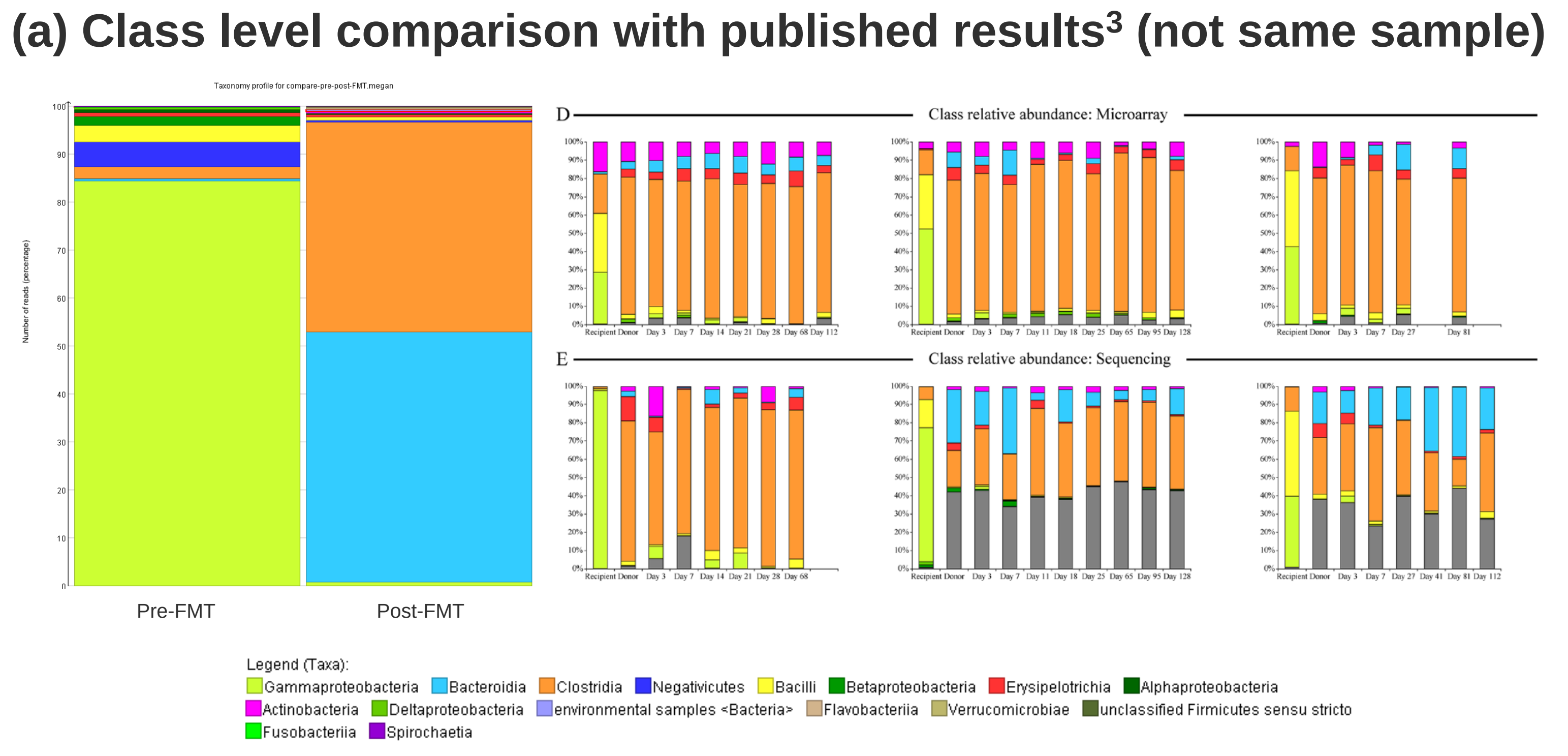
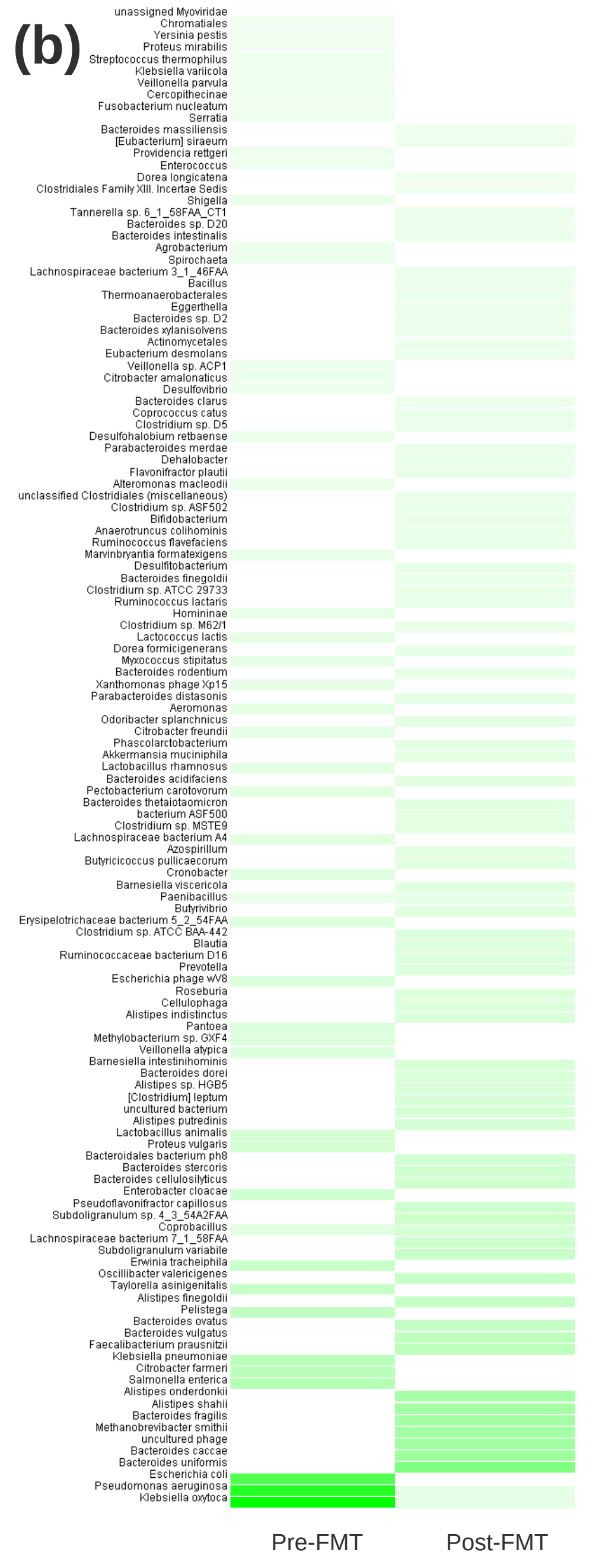


Figure 2. Species Histogram (A) and Taxonomic tree (B) showing abundance of community members in an infant stool sample using CCS profiling. Nodes of the tree are scaled to the number of assigned reads. Most reads can be assigned at the species or genus level.

Clostridium difficile Patient, Fecal Microbiota Transplant (FMT)





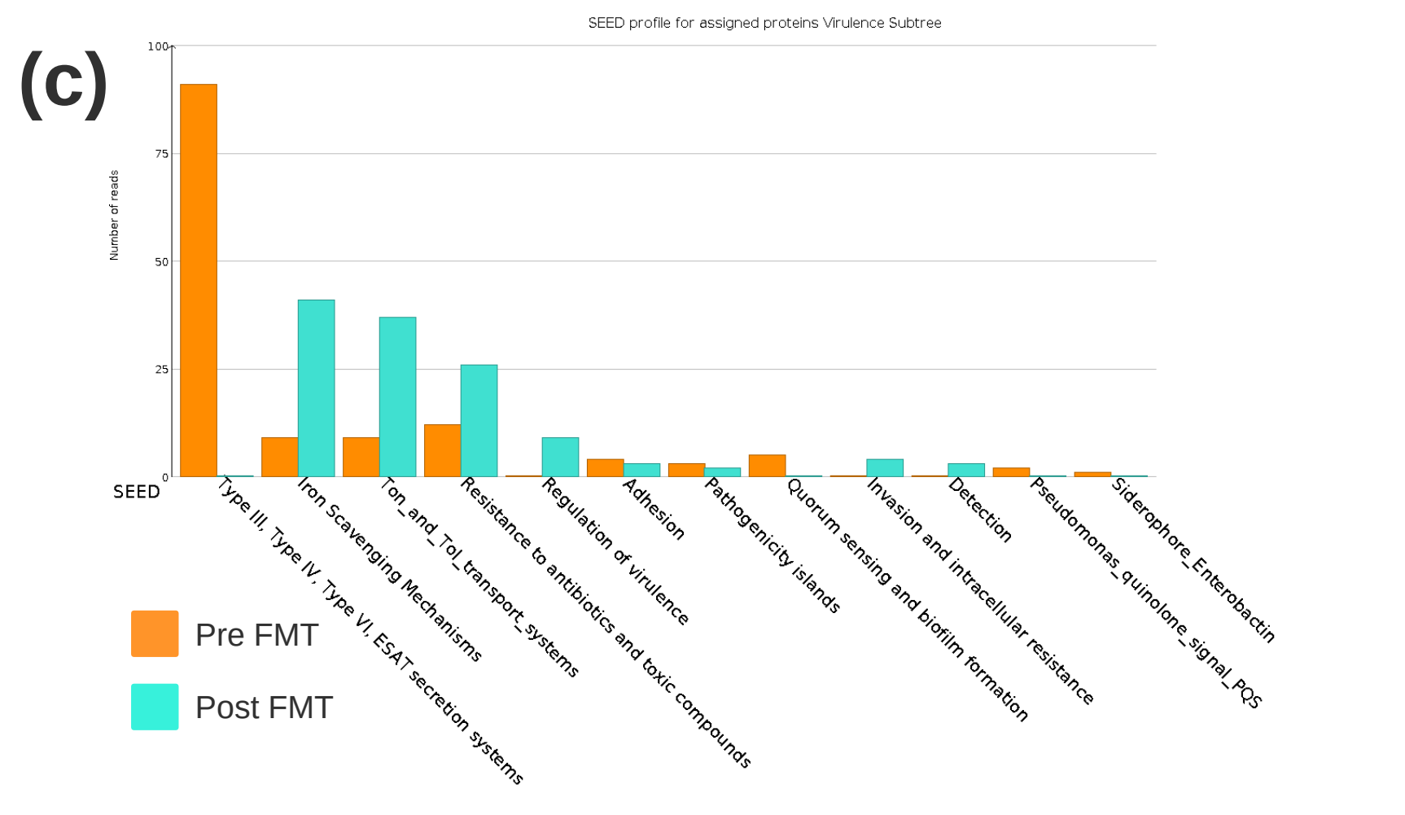


Figure 3. CCS profile analysis of pre- and post-fecal transplant samples. (a) Class level assignments compare well with previously published results from similar samples characterized using microarrays and 16S techniques. **(b)** Full species-level classification shows very distinct community profiles pre- and post-transplant. **(c)** Profiling using genes allows data to be visualized using SEED[5] functional classifications, here pre-FMT secretion system genes are seen that are not present post-FMT.

Conclusions

- Metagenomic CCS profiling offers distinct advantages over both 16S and shotgun assembly methods.
- While having a high tolerance for sample input problems such as low input quantities and fragmented DNA, CCS profiling allows species level taxonomic classification and functional studies.
- Using a mock community, we demonstrate abundance measurements comparable to 16S quantification.
- Using primary samples we show comparable results to both 16S and microarray data, while allowing finer grain species level classification and meaningful functional insight.

References

[1] Hyatt D, *et al.*, Gene and translation initiation site prediction in metagenomic sequences. Bioinformatics. 2012 Sep 1;28(17):2223-2230.

[2] Huson D.H. *et al.*, Integrative analysis of environmental sequences using MEGAN 4, Genome Res. 2011. 21:1552-1560.

[3] Shankar V. *et al*. Species and genus level resolution analysis of gut microbiota in *Clostridium difficile* patients following fecal microbiota transplantation. Microbiome. 2014 Apr 21;2:13.

[4] Quast C, *et al.*, (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucl. Acids Res. 41 (D1): D590-D596.

[5] Overbeek R, *et al.* The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. 2005 Oct 7;33(17):5691-702.