Metagenomics and Taxon Assignment: Literature Review

Presented by: Thomas Cowell

12-4-2018

What is Metagenomics?

- Environmental genomics: genomic analysis of microbes

- Why study metagenomics?
  - Understanding the diversity of life on Earth

  - Microbiome impacts human health

Approaches to Metagenomics

- Whole genome Shotgun Sequencing
  - Extract environmental DNA and randomly shear into short fragments
  - Read frequencies match population frequencies
  - Entire genomes are sampled with low depth

- Targeted metagenomics
  - PCR amplification of a specific target sequence
  - Amplification bias masks population frequencies
  - Taxonomically informative regions are sampled with high coverage

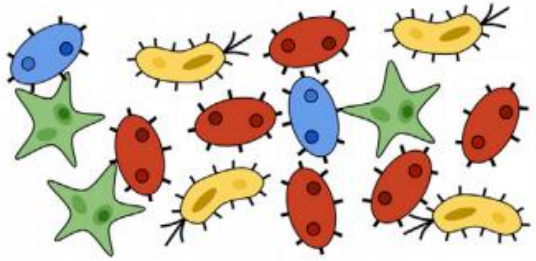
Environmental Metagenomics

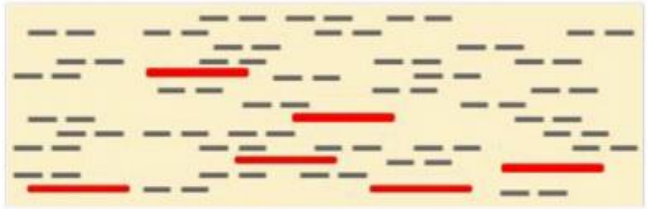
- Classify global bacterial diversity
  - Estimations of total diversity disagree dramatically (10^7 to 10^{12})
  - Identifying the total diversity in a given sample is challenging

- Typical experimental designs are not sensitive to rare species
  - Lack of amplification or PCR bias
  - Primer matching
  - Sensitivity of target sequence
  - Conserved sequences might be misleading

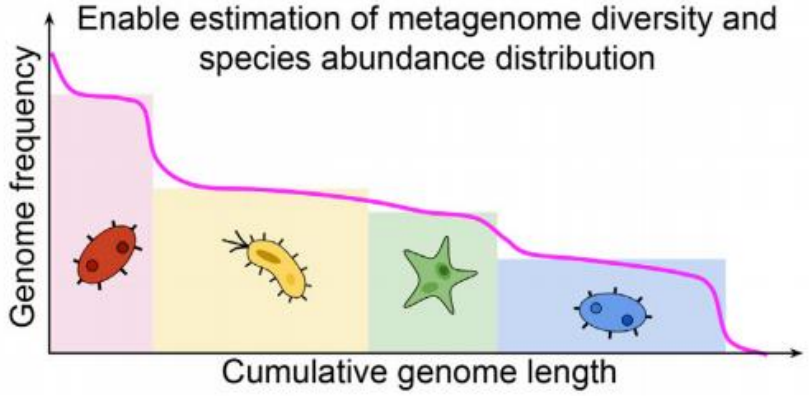
Long Reads Enable Accurate Estimates of Complexity of Metagenomes

Hybrid metagenome sequencing with short and long reads





And alignment between short and long reads



Long Reads Enable Accurate Estimates of Complexity of Metagenomes

- Combines many short reads with a smaller number of long reads
- Long reads act as templates for aligning and assigning short reads

- Developed an estimator of metagenomic capacity and abundance for rare species with low coverage

- Long reads are expensive, but improve the usefulness of short reads (~230 million short reads, ~50,000 long reads)

Taxon Assignment

- Identify the various species present in an environmental sample
- Understand the relative abundances of species

- Short fragments are often mapped to a species level reference genome

- Reference databases are not complete
  - Rare species are less likely to be represented in the reference
  - Rare reads are more likely to be mapped inaccurately

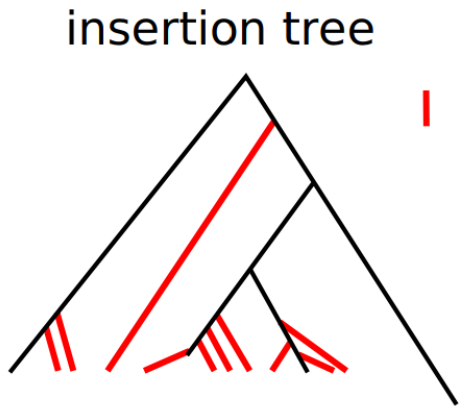
Accurate Reconstruction of Microbial Strains from Metagenomic Sequencing Using Representative Reference Genomes

- A new method for taxon assignment: Strain Prediction and Analysis using Representative SEquences (SPARSE)
- Hierarchical clustering on all genomes from a large reference database
- Compute a single representative sequence for each cluster
- Model the probability that a given read does not map to any genome in the reference

- SPARSE provides high taxonomic resolution

Phylogenetic Placement of Exact Amplicon Sequences Improves Associations with Clinical Information

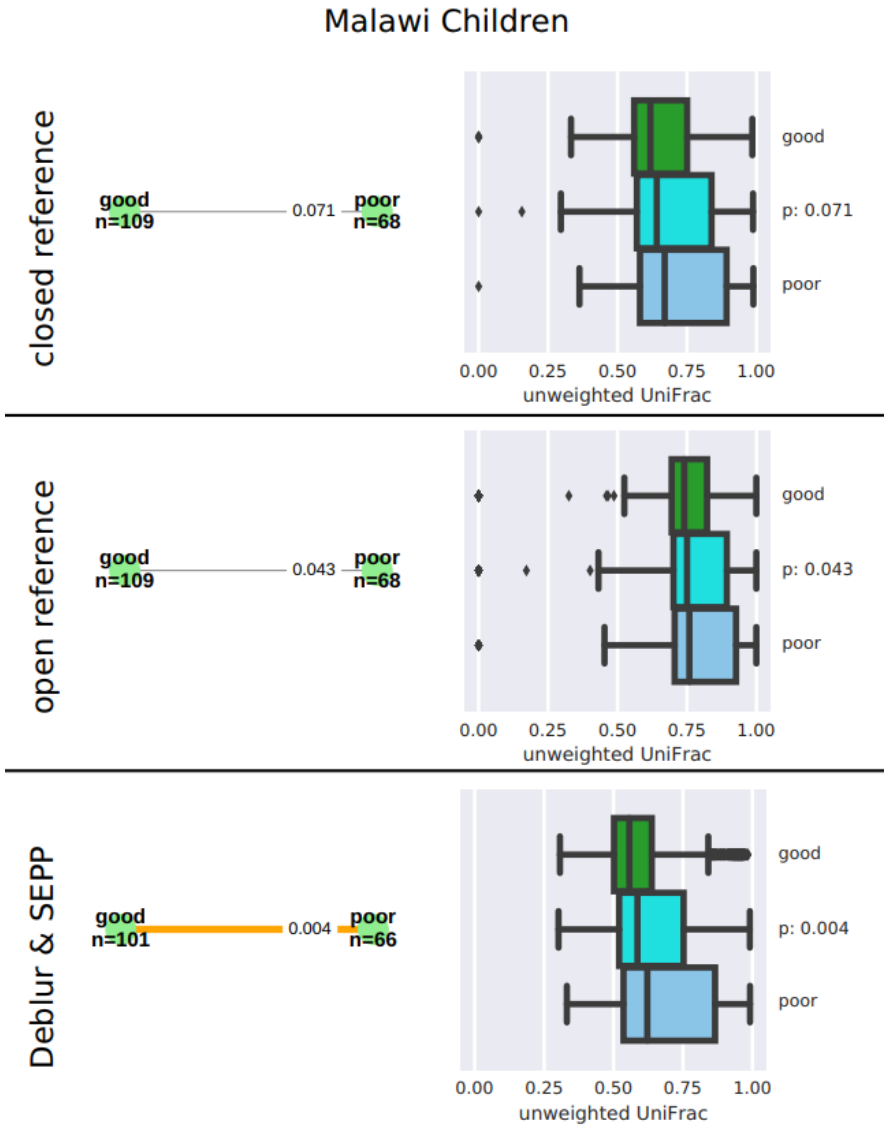
- Highly-scalable method for taxon assignment: SATé-Enabled Phylogenetic Placement (SEPP)
- Uses reference phylogeny to enhance accurate placement of short reads



pro: reference phylogeny
pro: keep most sOTUs

SEPP Distinguishes Clinical Variables

- SEPP distinguishes patient groups with the highest significance
- Higher taxonomic resolution relative to OTU based methods

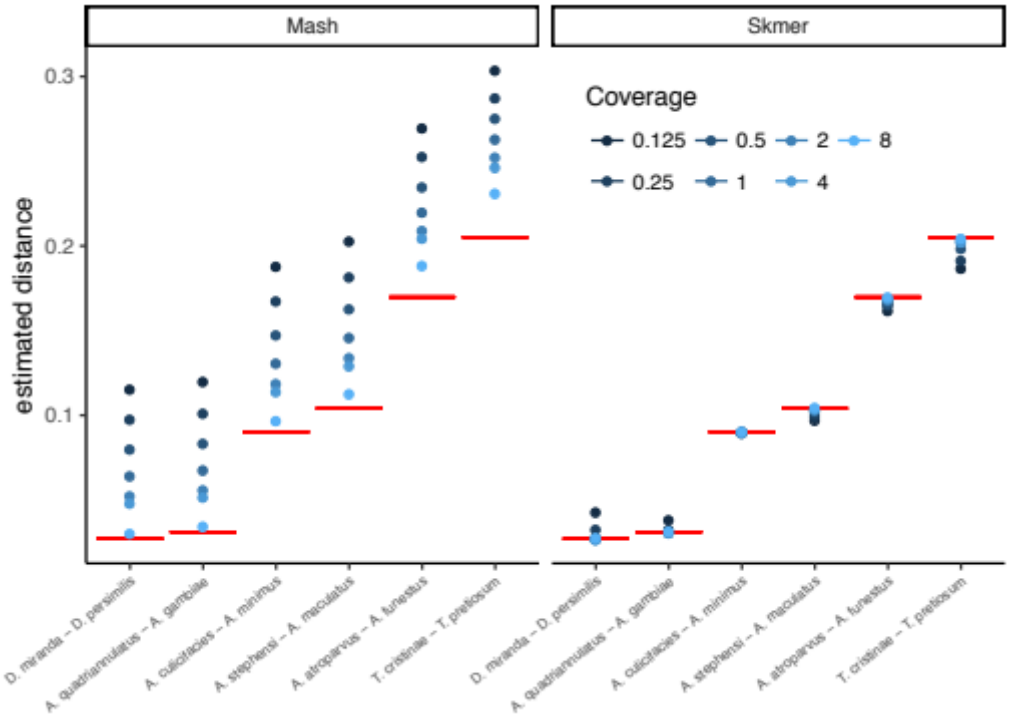


Assembly-free and alignment-free sample identification using genome skims

- Shotgun sequencing often heavily samples chloroplast and mitochondria derived DNA
- Traditional methods require genome assembly of the organelle

- Skmer: A method to use the entire set of unassembled reads as a “genome skim” and map it to the closest reference genome skim
- Computes the distance between two skims without assembly and with low coverage

Improved Taxon Assignment at Low Coverage



Open Problems

- Placement of the query skim in the reference phylogeny
- Decomposition of a genome skim into its constituent species

Future Work

- Read in-depth: important papers referenced by these works
- Compare each method:
  - scope of scientific questions addressed
  - strength of results
  - quality of analysis
  - potential for improvements
  - Limitations of each study