Whole Genome Metagenomics

Dr. Jung Soh

657.000 SS 2017

Whole genome metagenomics

- Advantages (over marker genes)
  - No primer (target region) bias
  - Improved resolution to identify lower taxonomic ranks
  - All microbes (eukaryotes, viruses) can be identified
  - Functional information is included
- Disadvantages
  - Host or site contamination
  - Bioinformatics challenge
- Goals
  - Taxonomy analysis
  - Functional analysis
  - Integrated analysis

Taxonomy analysis

- Relative abundance of different microbes in a sample or between samples
  - Taxonomic binning (into source organisms)
- Issues
  - Reads are from many unidentified organisms
  - Use reads or assembled contigs?
  - Limits of taxonomy database
  - Limits of taxonomic classification algorithms

Binning approaches

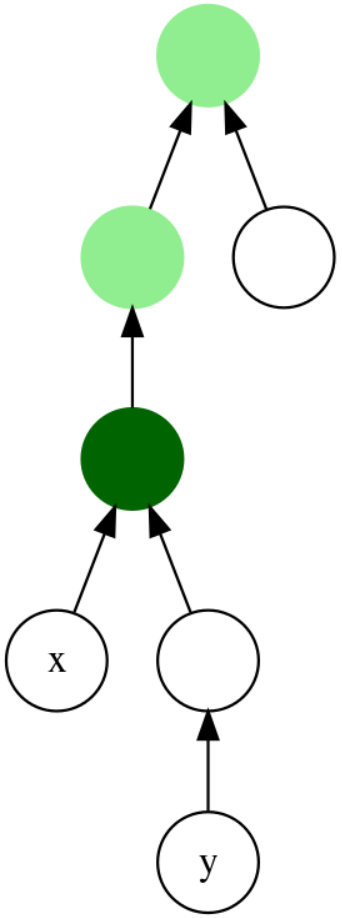
- Composition-based
  - Uses compositional features (GC %, *k*-mer frequency distributions)
  - Standard classifiers (Bayes)

- Similarity-based
  - Search for similarity in reference database (BLAST)
  - Assign reads to organism of best-hit reference or use an algorithm to determine the source organism

Lowest common ancestor (LCA)

- Commonly used taxonomic binning algorithm
  - Works best with BLAST results
  - Conservative assignment based on tree of taxa

- Two steps:
  - Collect a subset of BLAST hits above a threshold
    - Absolute bit score, % of best bit score
  - Assign sequence to the lowest-rank taxon that covers all taxa of those hits



https://en.wikipedia.org/wiki/Lowest_common_ancestor

Taxonomic classification tools

- *k*-mer-based DNA-level classification
  - Kraken
  - CLARK
- Read-based protein-level classification
  - Kaiju
- LCA algorithm
  - MEGAN

Functional analysis

- Map sequences to functional categories
  - Stress response
  - Amino acids and derivative
  - Nitrogen metabolism
  - Virulence factors

- Or to specific orthologs (if possible)
  - KEGG IDs
    - K0001 (alcohol dehydrogenase)
  - EC numbers
    - EC 1.1.1.1

Functional databases

- COG
  - Clusters of orthologous groups (NCBI)
- SEED
  - Used by RAST and MG-RAST systems
- Pfam
  - Protein domains
- EggNOG
  - 190K orthologous groups covering 2031 organisms
- KEGG
  - 12K orthologs (KOs), linked to modules and pathways
  - Popular, but full access is no longer free

Assembly in metagenomics

- Reads assembled as in *de novo* genome assembly
- Pros
  - Contigs are easier to process computationally than reads
  - Informative data when reads are too short
- Cons
  - Assembly is computationally intensive
  - Overly diverse low-coverage reads can fail to assemble
  - Chimeras can be easily formed
  - Reads from abundant organisms assemble better, leading to bias in analysis
  - Assembly quality hard to evaluate

Functional analysis tools

- Web-based
  - MG-RAST
  - IMG/M
  - EBI Metagenomics Server
- GUI-based
  - MEGAN (combined analysis of taxonomy and function)