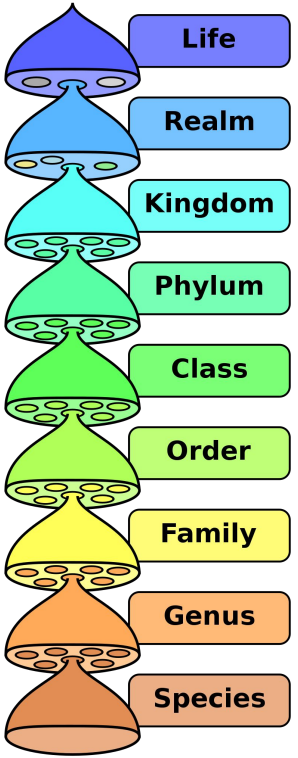
Taxonomic Classification with MiCoP

Tutorial #8

Karanveer Singh
Dan Drzewicki

Taxonomy

- Naming, grouping, and ranking of biological organisms
- Tree like structure
  - Layers of the tree represent taxonomic ranks
  - Genetically similar organisms are closer on the tree
    - Have an ancestor node in common
- Taxonomic ID
  - NCBI assigns a numeric ID for each node in taxonomy tree
  - Can trace back to find the lineage given an Tax ID



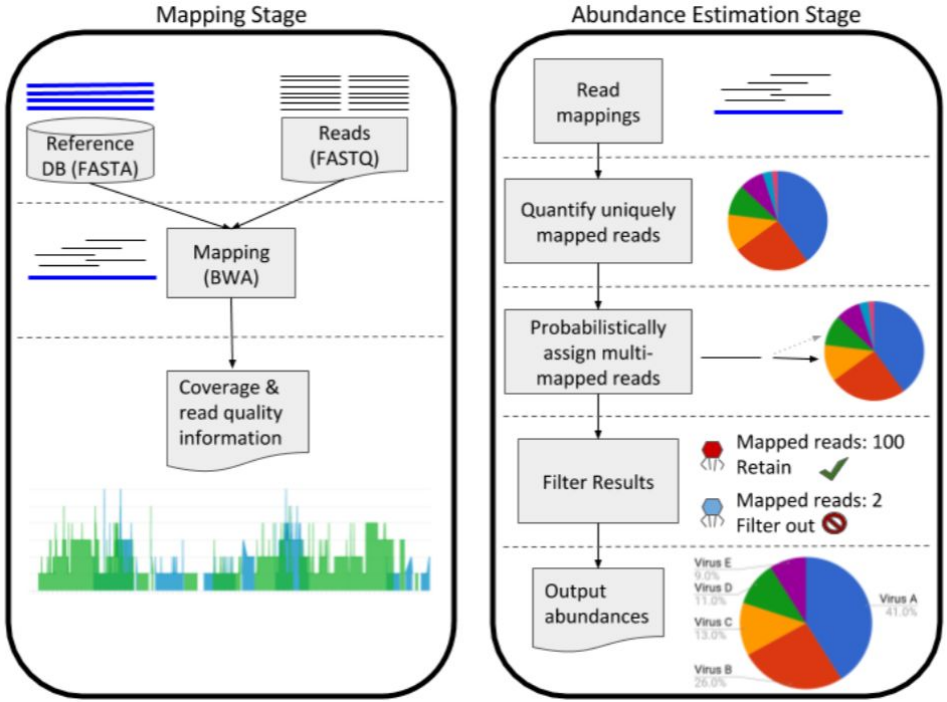
Microbial Communities

- Assemblage of interacting organisms that together form a community
- Microbial ecology
  - Studying the interactions of organisms amongst themselves and their environment
- Sequencing technology allows us to take reads from these communities
  - Question? : How do we know which organisms exist in a community?
    - Microbial community profiling (MiCoP)

Microbial Community Profiling

- Type of taxonomic classification
  - Taxonomic ID is assigned to the reads from the community
- Profiling
  - Find out who and how many
  - Find out which organisms exist in a community as well as their abundance
    - Relative abundance - how common is a species relative to others in the same community
- MiCoP
  - Method for calculating relative abundance of taxonomic levels within a community given a fasta/fastq file
  - Currently only works for viruses and fungi

MiCoP Pipeline



Map reads to reference database.

Two-step process to filter and profile results.

Reference Database and Mapping

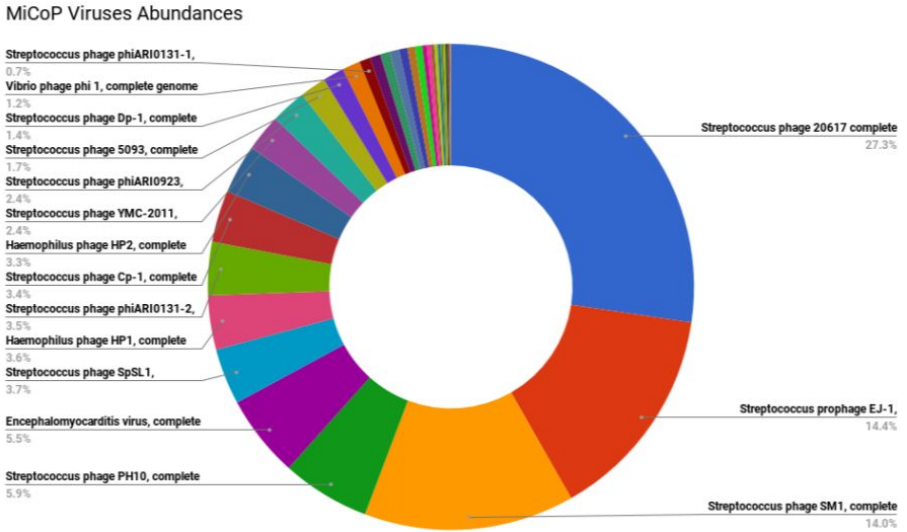
- Reference database dependent to a large extent on the database used
  - Smaller databases
    - Lower sensitivity
    - Fast to search
  - Larger databases
    - Longer to search
    - More accurate
- Due to increasingly powerful hardware and fast mapping algs, large databases can be used realistically
- MiCoP opts to use the full NCBI RefSeq Viral and Fungal databases

Probabilistic Assignment of Multi-Mapped Reads

- Two-stage process:
  - Uniquely mapped reads are immediately assigned
    - Trivial
  - Multi-mapped reads are probabilistically assigned
    - Proper assignment has major impact on results
  - BWA chose which genome to assign multi-mapped reads to
  - Multi-mapped reads are assigned to a genome with probability equal to the relative uniquely-mapped read counts for each of those genomes
- < 10 uniquely-mapped reads are filtered

Relative Abundance Estimation

- Normalize read counts for each genome by the length of the genome
- Normalize the adjusted counts of each genome by the sum of the adjusted counts
    - Species abundances sum up to 100%



Performance Metrics

TP = species present in a sample is correctly predicted

FP = predicted presence but not actually in a sample

FN = species was present but presence not predicted

- Precision
- Recall
- F1 Score
- L1 Error
  - Accuracy of relative abundances
  - $$L1\ Error = \sum_{i=1}^{S} |Predicted_i - Actual_i|$$

MiCoP-Kraken Comparison

MiCoP was evaluated using simulated data compared to a ground truth

- MiCoP shows order of magnitude improvement in abundance estimation

	L1 Error	Precision	Recall/Sensitivity	F1-Score
MiCoP	0.00909	1.0	1.0	1.0
Kraken	1.15466*	0.82222	0.925	0.87059

Viral community

	L1 Error	Precision	Recall/Sensitivity	F1-Score
MiCoP	0.09124	1.0	0.98155	0.99069
Kraken	1.15834*	0.85147	0.90959	0.87957

Fungal community

- MiCoP only capable of classifying these two communities
  - Kraken can classify “any” community
- Kraken is a read classification method, not a relative abundance estimation method
- MiCoP ultimately slower but more accurate than Kraken

Demo

miCoP Accuracy

Dataset	Strain	MiCoP predicted relative abundance (%)	Actual relative abundance (%)
Low Complexity Virus [1]	Olive latent virus	28.59	27.93
	Wheat eqlid mosaic virus	15.11	14.59
	Enterobacteria phage RB16	8.52	8.33
	Pseudomonas phage 73	7.63	7.43
High Complexity Virus [1]	Shigella phage pSb-1	3.31	2.22
	Propionibacterium phage P105	3.10	2.21
	Tobacco streak virus	0.00	1.87
	Prochlorococcus phage P-HM2	2.24	1.59

[1] Conceicao-Neto, N., Zeller, M., Lefrere, H., De Bruyn, P., Beller, L., Deboutte, W., ... & Matthijnssens, J (2015). Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. *Scientific reports*, 5, e16532. doi:10.1038/srep16532