Metagenomics

Ahmet Sacan

Slides adapted from:
https://bioinformatics.ca/workshops/2015/analysis-metagenomic-data-2015

Metagenomics

- Microbiome: Collection of microorganisms living in an environment
- Metagenome: Collective genomes of these microorganisms.
- Marker-gene based approaches (e.g., 16S metagenomics)

Metagenomics vs 16S

- Metagenomics: sequencing ALL the DNA in a sample
- 16S: sequencing 16S rRNA gene (or a portion of it)

- Metagenomics - Pros
  - Provides functional information: "What are they?"
    - Which genes? What biochemical functions?
  - Less bias from sequencing
  - Can identify all microbes (eukaryotes, viruses, etc.)
- Metagenomics - Cons
  - Host/site contamination can be significant
  - Expensive (more sequencing depth is required)
  - May not be able to sequence “rare” microbes
  - Complex bioinformatics

Metagenomics Analysis

- Assemble reads into contigs (?)
- Assign reads/contigs to:
  - organisms (OTUs)
  - genes
- Relative abundance of organisms
- Functional annotation of gene sets

Read assignment

- Similar to RNA-Seq experiments, but now need to consider all (or many) species instead of a single experimental species.
- Binning approaches
  - Assign each read to an OTU by similarity search or statistics
- Marker approaches
  - Pre-determine representative marker genes for OTUs
  - Assign reads only to these markers.

Binning

- Composition-based
  - Uses GC composition or k-mers statistics
  - Very fast
  - Generally not very precise and not recommended
- Sequence-based
  - Compare reads to large reference database (e.g., BLAST)
  - Reads are assigned based on "Best-hit" or "Lowest Common Ancestor" approach
  - Examples: MEGAN, MG-RAST, Kraken

Binning vs Marker

- Binning
  - May be too computationally intensive
  - May not adequately reflect organism abundances due to genome size
- Marker
  - Doesn't allow functions to be linked directly to organisms
  - Genome reconstruction is not possible
  - Very sensitive to choice of markers

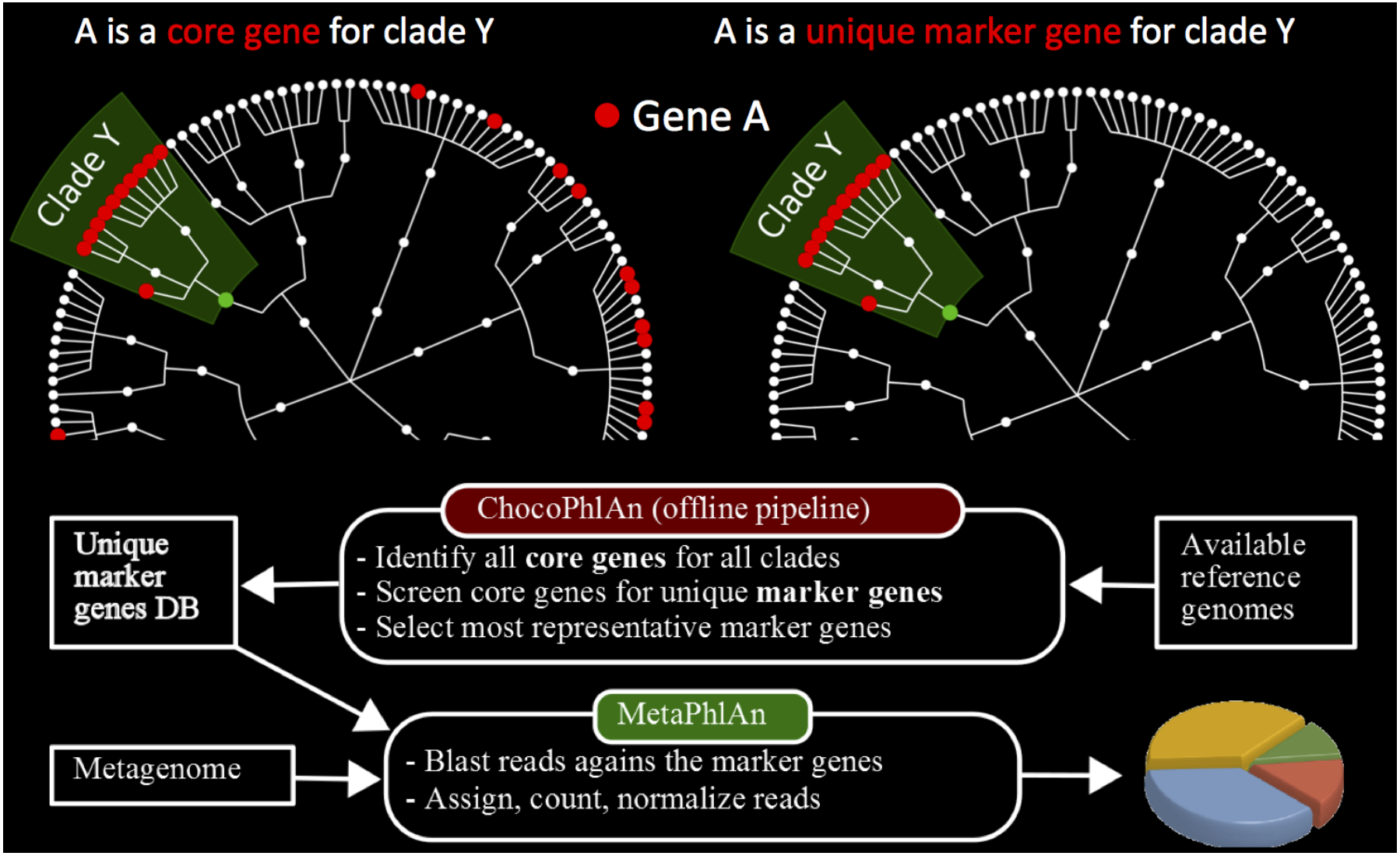
Marker

- Single Gene
  - Identify and extract reads hitting a single marker gene (e.g. 16S, cpn60, or other “universal” genes)
  - QIIME, Mothur
- Multiple Gene
  - Several universal genes
    - PhyloSift (Darling et al, 2014)
      - » Uses 37 universal single-copy genes
  - Clade specific markers
    - MetaPhlAn (Segata et al, 2012)

MetaPhlAn

- Uses "clade-specific" gene markers
- A clade represents a set of genomes that can be as broad as a phylum or as specific as a species
- Uses ~1 million markers derived from 17,000 genomes
  - ~13,500 bacterial and archaeal, ~3,500 viral, and ~110 eukaryotic
- Can identify down to the species level (and possibly even strain level)
- Can handle millions of reads on a standard computer within a few minutes
  - Bowtie for read assignment

MetaPhlAn


A is a core gene for clade Y
A is a unique marker gene for clade Y
Gene A
Clade Y
Clade Y
ChocoPhlAn (offline pipeline)
- Identify all core genes for all clades
- Screen core genes for unique marker genes
- Select most representative marker genes
Unique marker genes DB
Available reference genomes
MetaPhlAn
- Blast reads agains the marker genes
- Assign, count, normalize reads
Metagenome