

# **Metagenomic Analysis**

**Dr. Jung Soh  
MOL.923 SS 2018**

# Metagenomics

- ▶ Deals with samples taken directly from the environment
  - Soil, water, hot spring, oil sands, human gut, stool
  - Also called environmental genomics
- ▶ More complicated than genomics
  - Less control
  - Mixture of many organisms

# Metagenomics questions

- ▶ Who's there?
  - Taxonomic classification
- ▶ How many/much of them are there?
  - Community diversity analysis
- ▶ What do they do?
  - Functional analysis

# Comparative metagenomic analysis

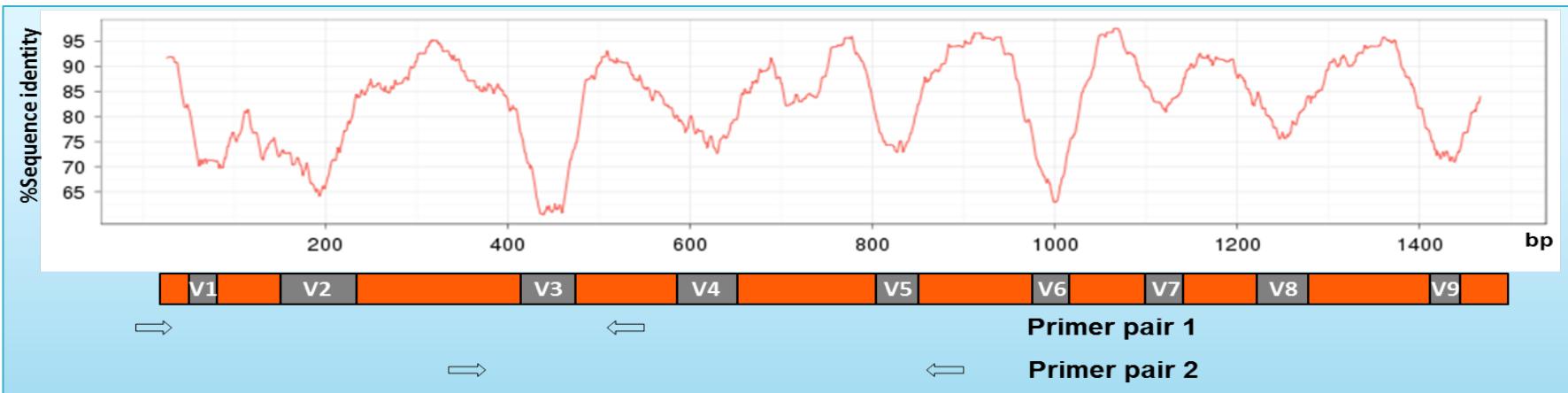
- ▶ Comparison across different environments
  - Individual taxon abundance
  - Overall microbial community composition
  - Dominant functions
  - Environment-specific functions of same taxon
- ▶ Within same environment
  - Changes over time in taxonomic composition and functions

# Two types of metagenomics

- ▶ Marker gene metagenomics
  - Targeted sequencing of one or more marker genes
  - Mostly bacteria can be identified
  - No functional analysis
- ▶ Whole genome metagenomics
  - Sequence whole metagenomes
  - All microbes can be identified
  - Functional analysis possible

# Marker gene metagenomics

- ▶ Use genes highly conserved yet unique to individual (mostly bacterial) species
  - Variable and conserved regions
  - 16S rRNA most frequently used
- ▶ Targeted sequencing using primer pairs



# Operational Taxonomic Units (OTUs)

- ▶ Computational objects for taxonomy analysis
  - 16S rRNA reads are clustered into OTUs
  - Usually at 97% sequence identity
- ▶ OTUs are mapped to a taxon
  - In theory, one OTU maps to one species
  - In practice, several OTUs map to one genus or species, or to higher taxonomic ranks



OTU	Taxon
OTU1	Rhzobiales
OTU2	Firmicutes
OTU3	E. coli
OTU4	E. coli
...	...
OTU 432	Bacteria

# Whole genome metagenomics

- ▶ Whole genome shotgun sequencing
  - Metagenomic reads come from many unknown organisms
- ▶ Reads assembled as in *de novo* genome assembly
- ▶ Metagenomic assembly challenges
  - Chimeric reads (single read from multiple organisms)
  - Severe coverage variation
- ▶ Assembly quality hard to evaluate

# Metagenomics resources

- ▶ MG-RAST (Metagenomic Analysis Server)
  - Web-based service
  - Usually overloaded, slow
- ▶ IMG (Integrated Microbial Genomes)
  - Databanks
- ▶ SILVA
  - Ribosomal RNA databases and tools

# Major metagenomics tools

- ▶ Qiime (Quantitative Insights Into Microbial Ecology)
  - Consists of Python scripts
  - Taxonomy and diversity statistics/visualization
- ▶ Mothur
  - Commands written in C++ programs
  - Taxonomy and diversity statistics/visualization
- ▶ MEGAN (Metagenome Analyzer)
  - Functional as well as taxonomy analysis
  - GUI with tree-based visualization

# Taxonomic classification tools

- ▶ Classifies metagenomic reads directly into taxa
  - Skips OTU clustering or contig assembly
  - Also called binning
- ▶ *k*-mer-based DNA-level classification
  - Kraken
  - CLARK
- ▶ Read-based protein-level classification
  - Kaiju

# Lab overview

