Marker Gene Metagenomics

Dr. Jung Soh

657.000 SS 2017

Considerations for marker genes

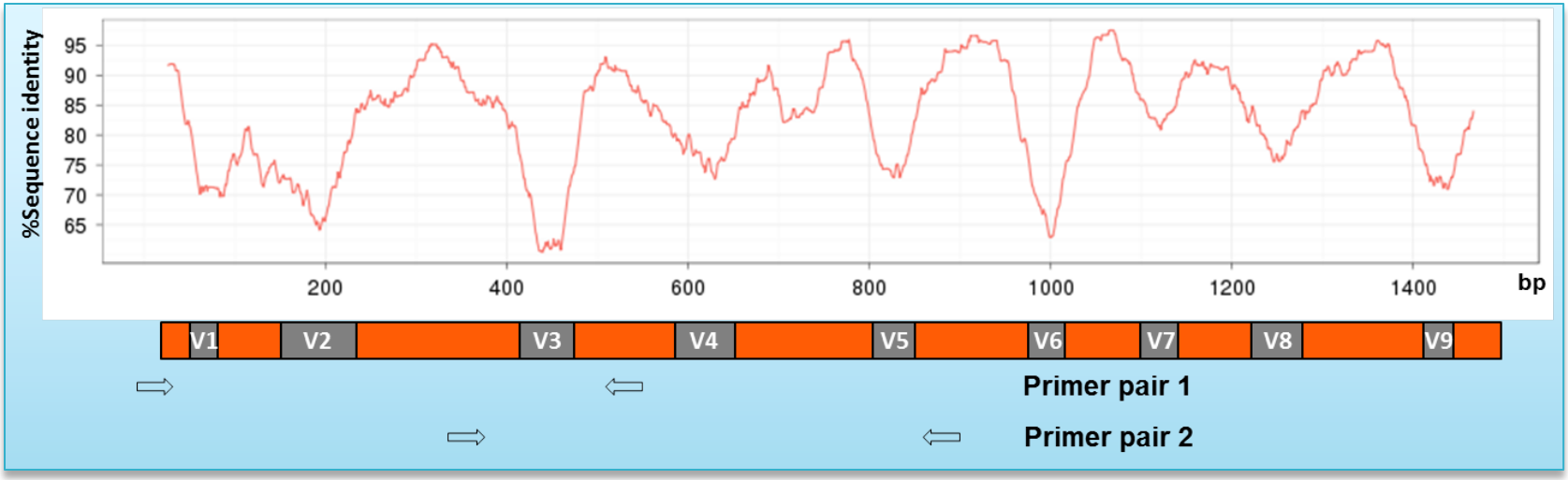
- Sufficient resolution to differentiate among communities of interest
  - 16S rRNA mostly down to genera and some species
- A reference database
  - Taxonomic assignment
  - Binning
- Standardization to enable comparison
  - Different targeted variable regions of 16S gene
  - Sampling protocols
  - Downstream bioinformatics analysis

rRNAs as markers

- Exist in all living organisms
- Play important roles in protein translation
- Can be used as molecular markers
  - For phylogenetic analysis
  - Used to build tree-of-life
- 16S rRNA most commonly used
  - Conserved, but sufficiently different across organisms

16S rRNA genes

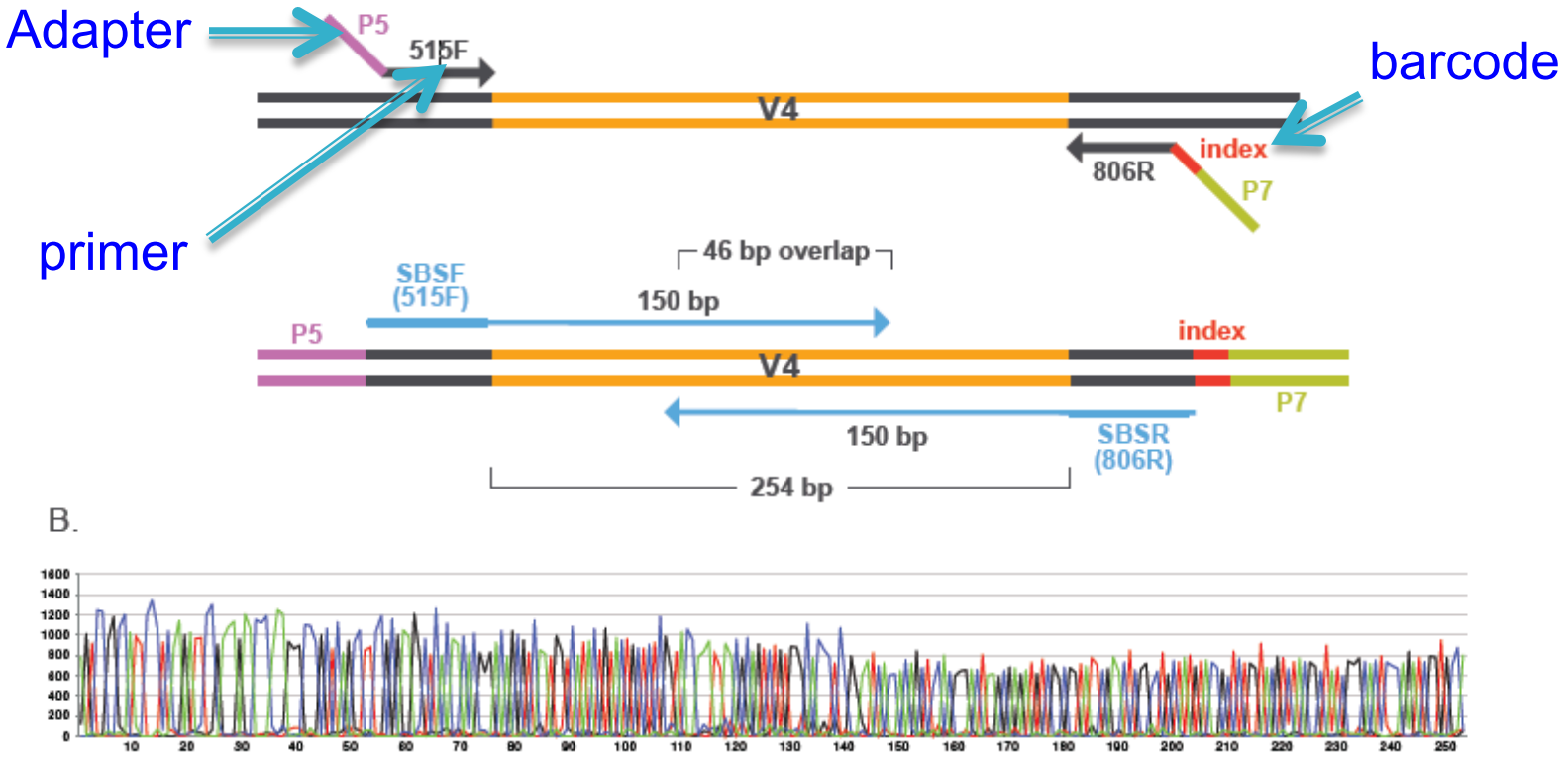
- Highly conserved yet sufficiently unique to individual (mostly bacterial) species
  - Consists of conserved regions and 9 hypervariable regions
    - V1, V2, …, V9
- Targeted sequencing using primer pairs



http://www.gatc-biotech.com/index.php?id=1025&L=1

16S target region amplification

Figure 1: Amplification Strategy and Perfect Paired-End Read



A. V4 was amplified from each sample using primers 515F and 806R tailed with P5 and P7 sequences, respectively. Paired 150 bp sequencing gives a full-length 254 bp fragment of V4 with a 46 bp overlap. B. Raw intensities (matrix and phasing corrected) for an example perfect 254 bp paired-end read from the V4 library.

Source: Illumina Application Note

Operational Taxonomic Units (OTUs)

- Computational units for taxonomy analysis
  - Does not reflect real world!
- 16S rRNA reads are clustered into OTUs
  - Usually at 97% sequence identity
- Existing tools differ a lot in how they form OTUs
  - Major computationally demanding step
  - Different assumptions and heuristics

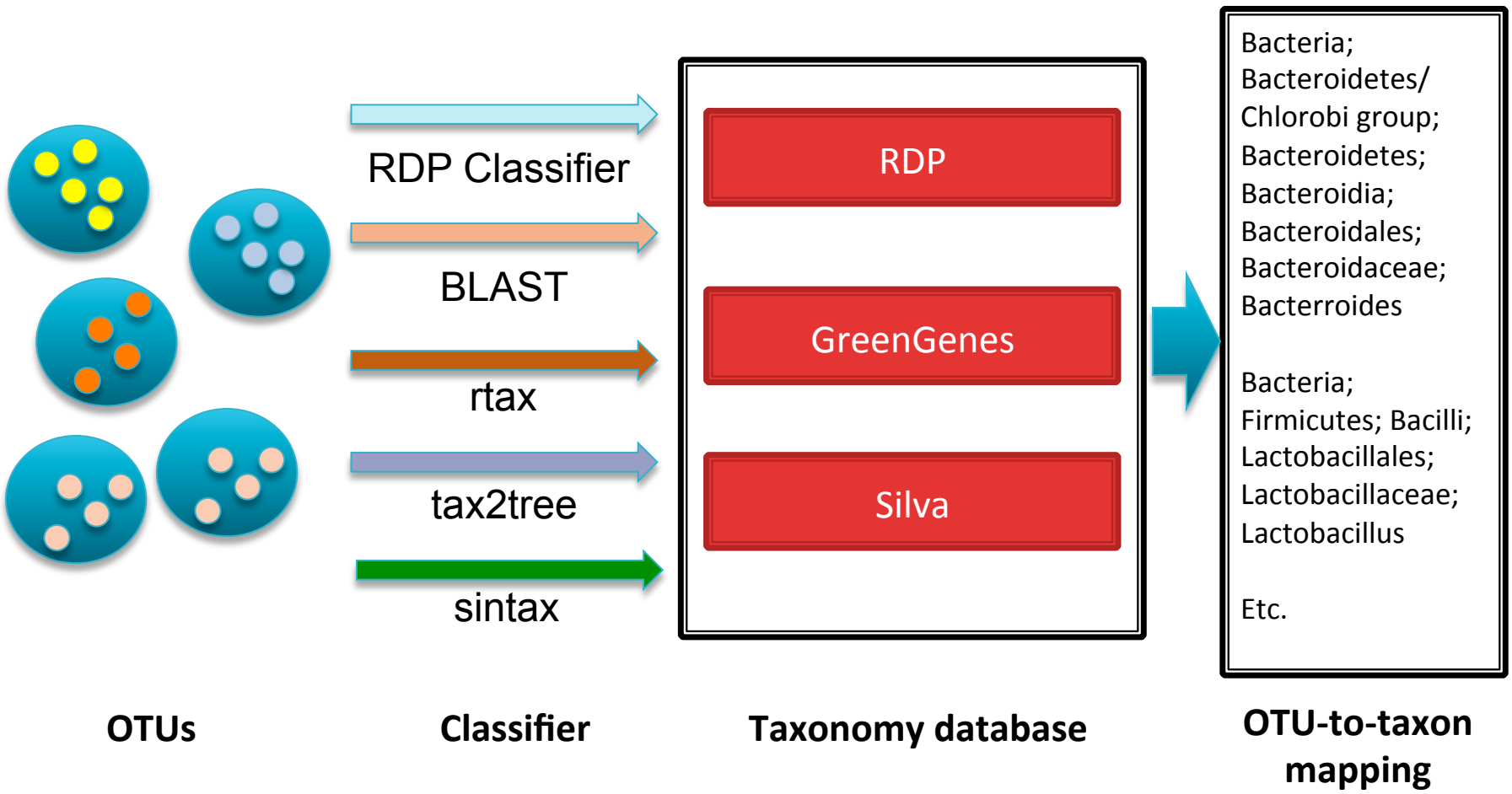
OTUs as new set of sequences

- One OTU is one (representative) sequence
- Reads belong to an OTU
- All downstream analyses performed on OTUs
  - OTUs should be as realistic as possible
- Selecting OTU (representative) sequence using member reads
  - Most abundant (frequently seen)
  - Centroid
  - Longest
  - First
  - Random

Taxonomic assignment

- OTU names are meaningless
  - otu1, otu2, …, otu4785
- We like meaningful names (annotation)
  - E. coli, Lactobacillus, …, Gammaprotebacteria
- OTUs are mapped to taxa
  - Theory (wish?): 1-to-1 (species) mapping
  - Practice: many-to-1 mapping
    - Mostly to genus (occasionally species) or higher taxonomic ranks

Taxonomic assignment of OTUs



https://bioinformatics.ca/metagenomics2016module-2-ppt

Taxonomy databases

- RDP
  - Most similar to NCBI Taxonomy
  - Relatively fast RDP Classifier is also provided
- GreenGenes
  - Preferred by QIIME
- SILVA
  - Preferred by mothur

OTU table

- Main result from OTU clustering
  - Each cell shows the number of reads in a sample in an OUT
  - All you need for diversity analyses
- OTUs can be taxon-mapped
- Rare OTUs can be filtered
  - May be due to sequencing errors and chimeras

	sample1	sample2	sample3	sample4	sample5
otu1	10	14	33	3	2
otu2	24	12	20	0	19
otu3	3	55	0	7	15

Alpha diversity analysis

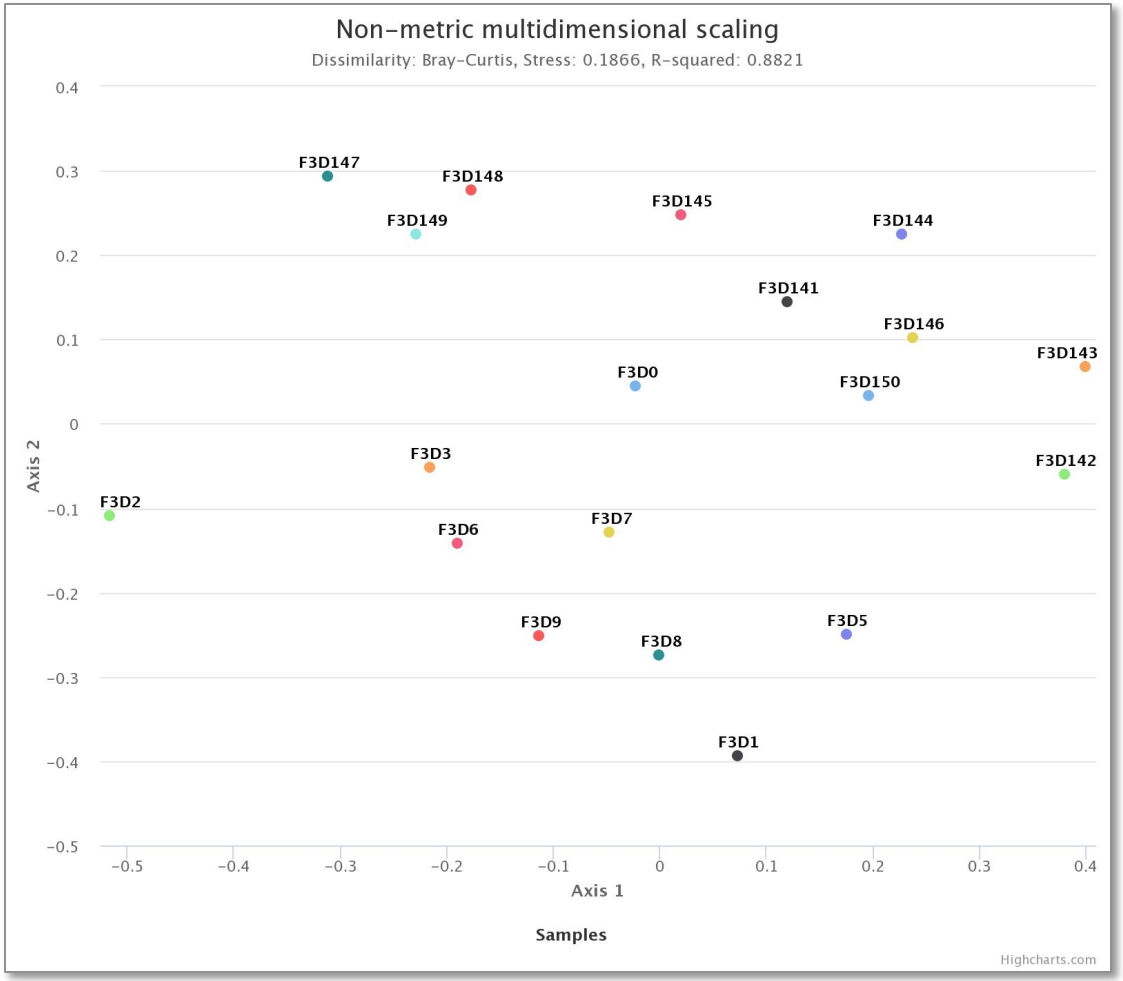
- Alpha diversity
  - Diversity of organisms in a single sample or environment
  - OTUs are assumed to be organisms (can be misleading)
- Two aspects of diversity
  - Richness: Number of organisms observed or estimated
  - Evenness: How evenly abundant are the organisms?
- Terms and metrics are used inconsistently!

Beta diversity analysis

- Beta diversity
  - Similarity or dissimilarity between samples or environments based on organism diversity
- Common measures
  - Bray-Curtis dissimilarity
    - OTU abundance used
  - Jaccard similarity:
    - OTU presence/absence only used
  - UniFrac distance
    - Incorporates a phylogenetic tree of OTUs

Ordination plots

- 2D visualization of beta diversity
- Most common
  - PCoA (Principal coordinates analysis)
  - NMDS (Non-metric multidimensional scaling)



Marker gene metagenomics tools

- Qiime
  - Python scripts (several dependencies, installation not simple)
  - Steep learning curve
  - Taxonomy and diversity statistics/visualization
- mothur
  - Collection of commands written in C++ programs (simple to install)
  - Can be slow
  - Taxonomy and diversity statistics/visualization
- USEARCH
  - Command line, single file to copy (no real installation)
  - Fast, no visualization
  - 32 bit free (max 4GB main memory), 64 bit paid, VSEARCH is 64 bit free (incomplete) alternative