Metagenomics introduction

Adam R. Rivers
Metagenome Program Lead
DOE Joint Genome Institute
February 4, 2016

Objectives

- Understand the main methods used in metagenomics
- Understand the main analysis techniques for each data type
- Understand the questions that metagenomics answers

Metagenomics defined

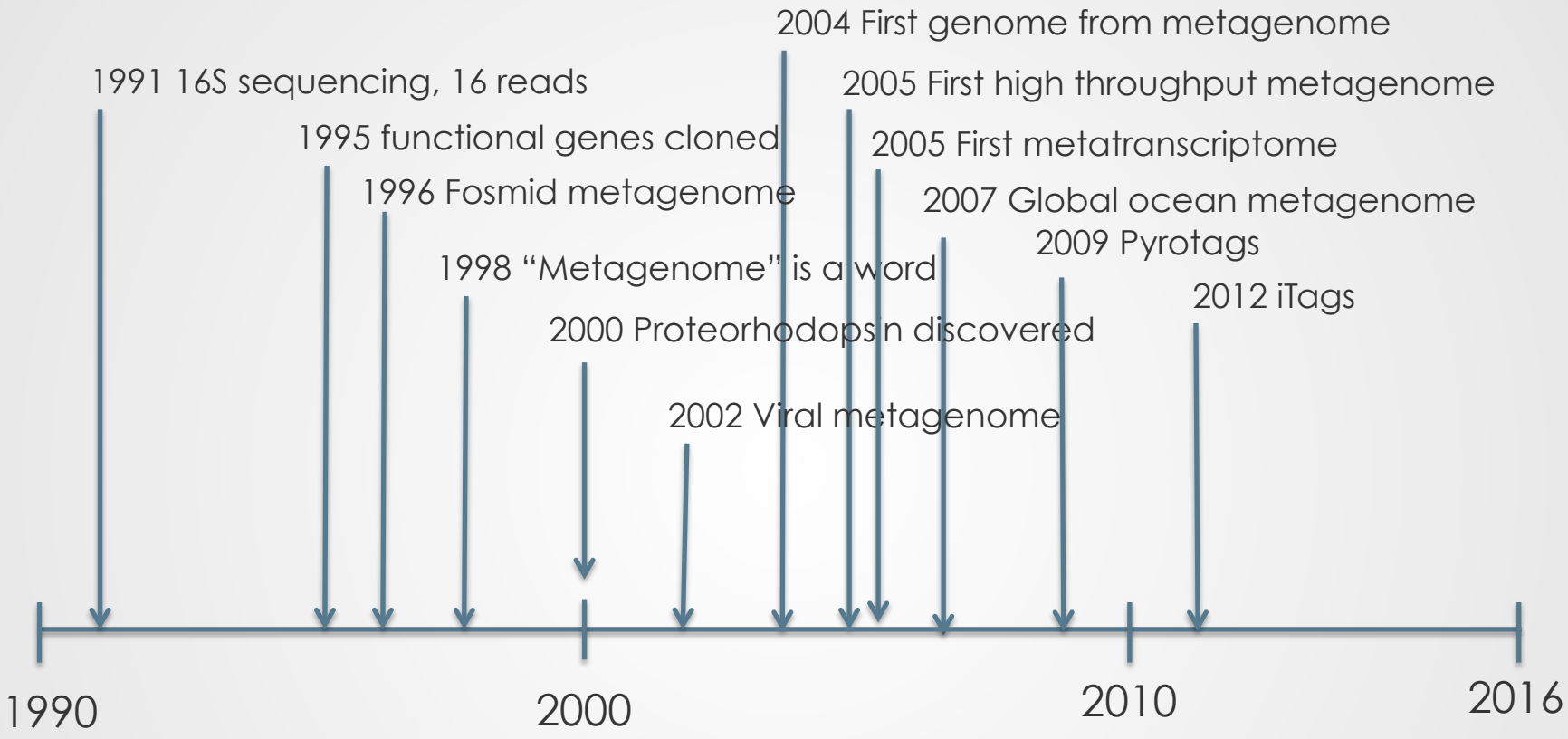
Core methods

- Amplicon sequencing
  - Phylogenetic markers
  - Functional genes
- Metagenomes
- Metatranscriptomes

Related methods

- Exo-proteomics
- Exo-metabolomics
- Single cell genomics
- Stable isotope probing

Metagenomics history

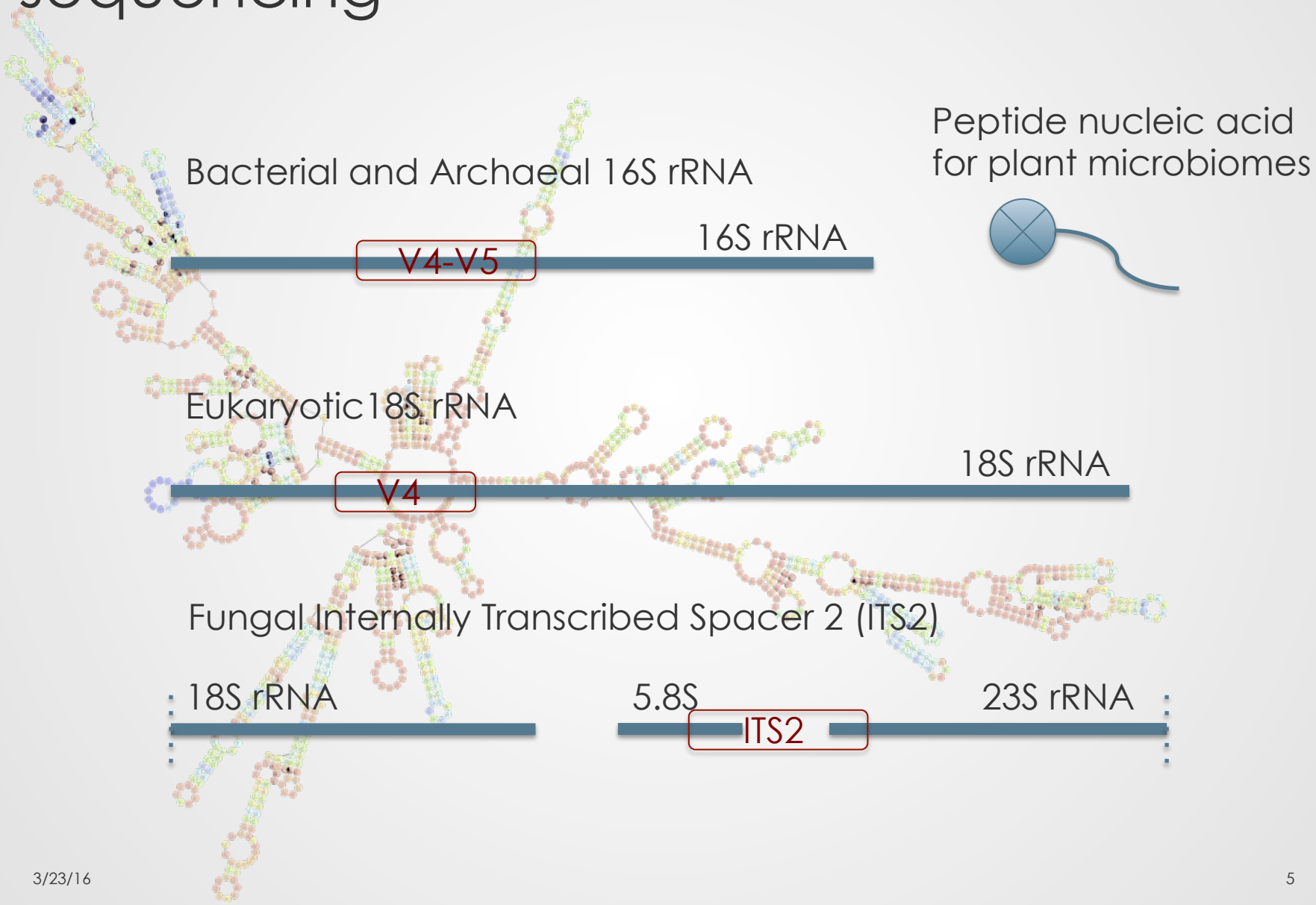


JOURNAL OF BACTERIOLOGY, July 1991, p. 4371–4378
0021-9193/91/144371-08$02.00/0
Copyright © 1991, American Society for Microbiology
Vol. 173, No. 14

Analysis of a Marine Picoplankton Community by 16S rRNA Gene Cloning and Sequencing

THOMAS M. SCHMIDT,† EDWARD F. DeLONG,‡ AND NORMAN R. PACE*
Department of Biology and Institute for Molecular and Cellular Biology, Indiana University, Bloomington, Indiana 47405

Main methods: 1. Amplicon sequencing



Peptide nucleic acid for plant microbiomes

Main methods: 2. Metagenomes

- A Metagenome is a collection of sequences from DNA in an environmental sample containing a mixture of organisms.

Taxonomic questions

"Who lives here?"

Functional questions

"What do they do?"

Organismal questions

"What of the genomes of the organisms in my community look like?"

Ecological and biogeochemical questions

Metagenome analysis

The first wave
2007-2013

- Read based bulk taxonomy
- Read based bulk functional annotation
- Fragment recruitment mapping to references

The second wave
2010-2014

- Assembly of metagenomes
- Mapping back to assembly
- Functions within taxonomy

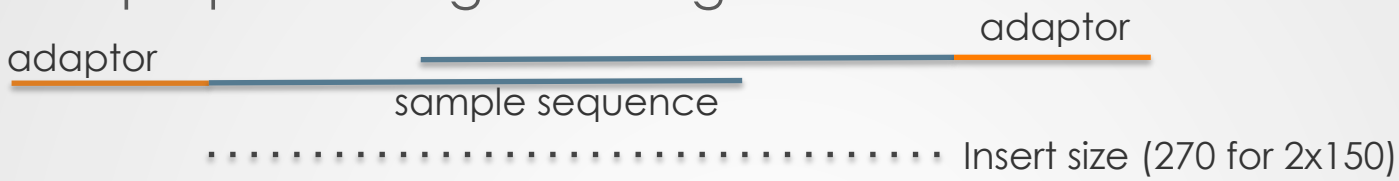
The third wave
2013-now

- Assembly and binning of genomes from metagenomes
- Metabolic potential within genomes
- Syntrophy
- Biogeochemical modeling

Metagenome example

Metagenome demonstration

- Data preprocessing cleaning



- Data file is in FASTQ

```
@MISEQ08:359:000000000-ALD3J:1:1101:8993:3210 1:N:0:CGCTCATGGCTCTG
GTCCTATTTTGGCCACCGGAAAATGTTCGGGATTTTTCGGTTTTGTACCGGGAAGGTTCTAGAAGGTTCCGAAGT
?AAAAFFFF3FDFGCFCEE0AEFDAB3F0AE0AFDDFG//A/EF0BB2F//E??A?1BGHGH2BGCGHHGEFGE?
@MISEQ08:359:000000000-ALD3J:1:1101:8993:3210 2:N:0:CGCTCATGGCTCTG
GGAGAGAATCCAGCAGCACAACGGCGTGGTGGTGGAAGCAGCGGGGATCTCGGCAGGTCTTCGCCCAGCTTCGCG
1>>11>>11DFF1BCA111ABAEEECEE///BE//BB0/B00AEEEEFE12FE>/EG0@BBFGEG/E>AG#####
```

- The purpose of data cleaning
  - Remove sequencing adaptors
  - Remove contaminant reads

Metagenome demonstration

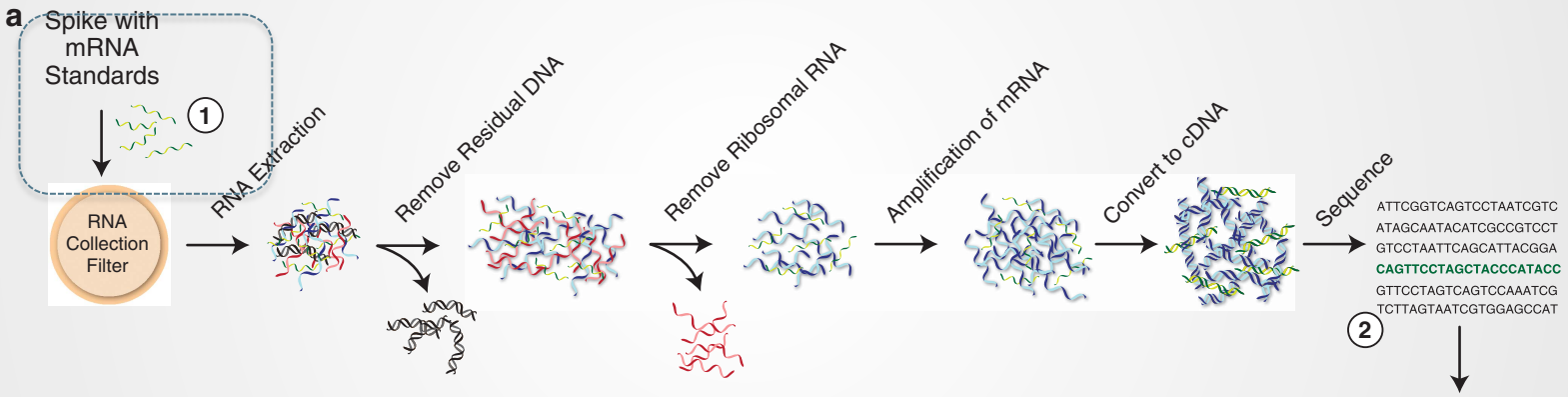
- Run once to remove primers
- `bin/bbduk.sh  in=data/ ref= data/ out=results/`

- Run again to remove contaminants
- `bin/bbduk.sh  in=data/ ref= data/ out=results/`

- Merge data together
- `bin/bbmege.sh  in=data/ out=results/mereged outu= results/unmerged`

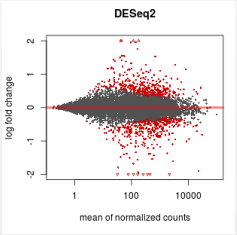
- Run a metagenome assembler and evaluate

Main methods: 3 Metatranscriptomes



Metatranscriptomes identify the actively transcribing community at a timescale of minutes to hours.

Typical metatranscriptome analysis

RNA sequencing	GATTCATA CCATTGG AATTCCCA	Illumina 2x150
Assembly	GATTAATA AATATTCATA CATATTCTA	Megahit
Annotation of genes		IMG
Mapping reads to assemblies		BBMap
Quantification		IMG
Differential expression		DESeq2, EdgeR, BaySeq

Thanks