16S Amplicon Sequencing Data Analysis:

the approach

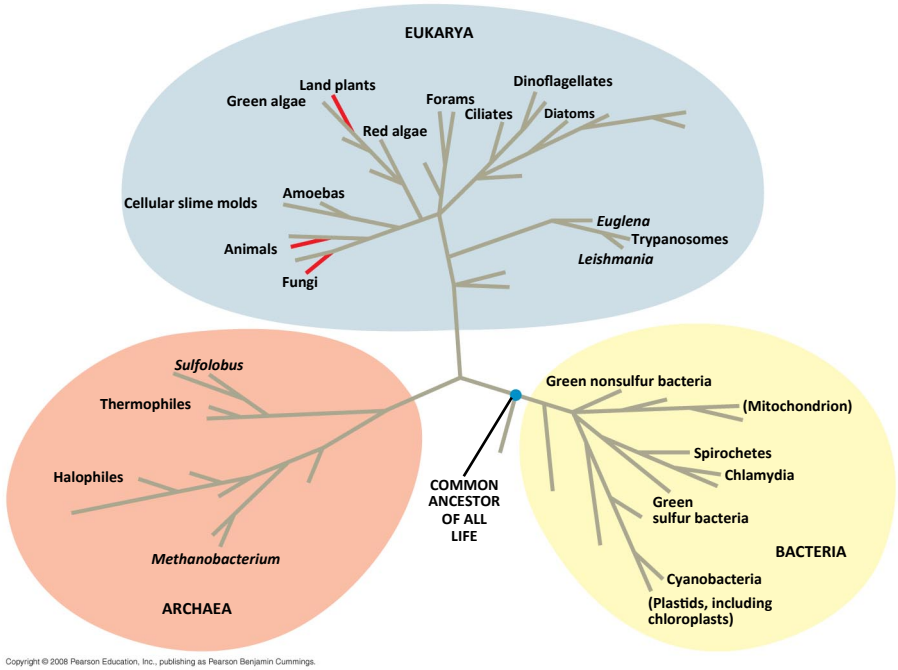
+

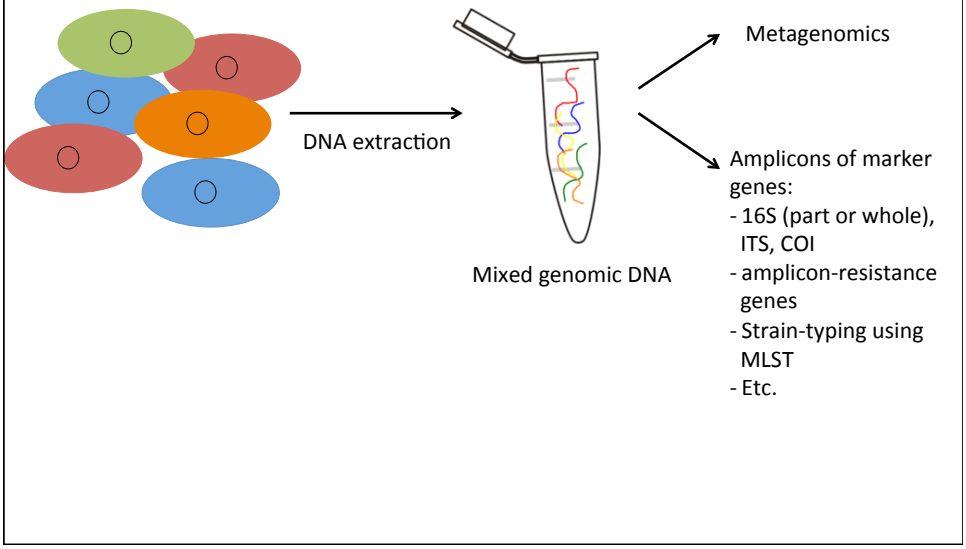
an example dataset

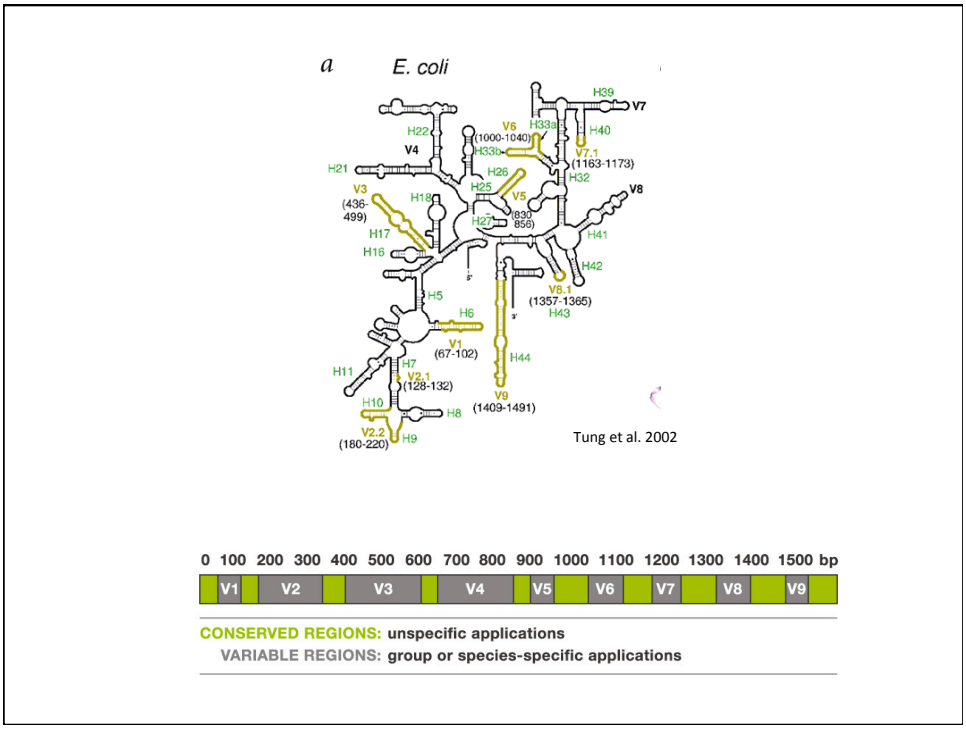
Rachel Adams

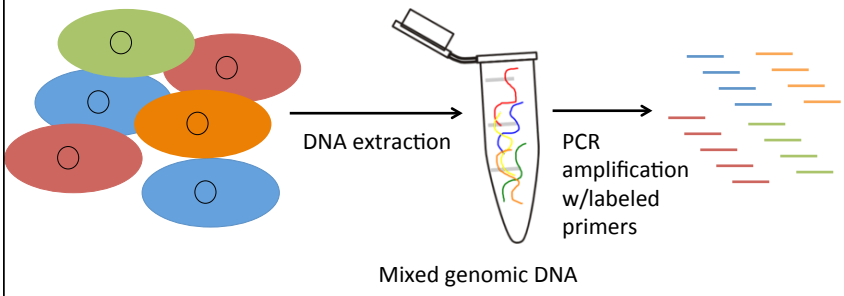
UC Berkeley - CGRL

December 4, 2014



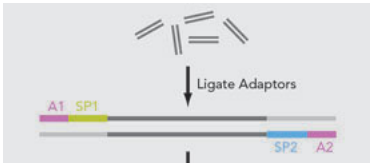

DNA extraction
Mixed genomic DNA
Metagenomics
Amplicons of marker genes:
- 16S (part or whole), ITS, COI
- amplicon-resistance genes
- Strain-typing using MLST
- Etc.


a
E. coli
Tung et al. 2002
0 100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 bp
V1 V2 V3 V4 V5 V6 V7 V8 V9
CONSERVED REGIONS: unspecific applications
VARIABLE REGIONS: group or species-specific applications

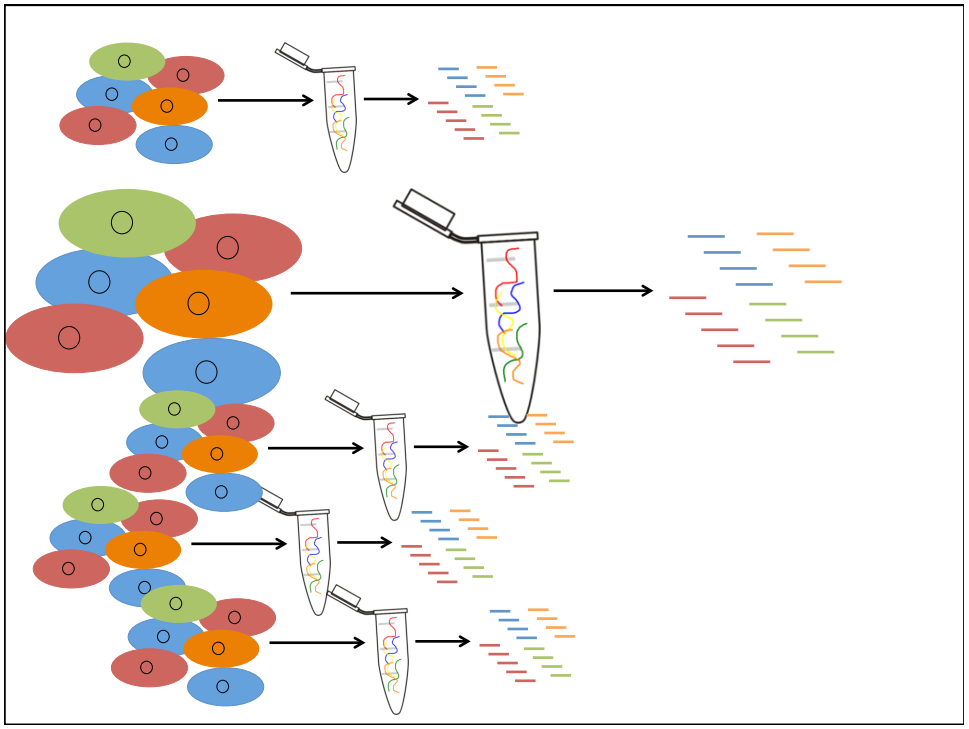


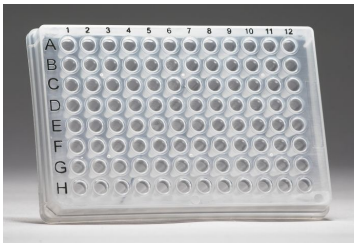
Mixed genomic DNA

Adapter1 +
Barcode +
Read1 +
Gene primer



Adapter2 +
Read2 +
Gene primer

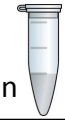




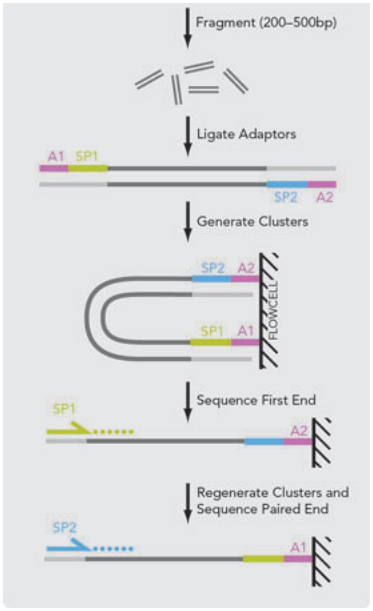
Clean-up PCR products

Determine concentration

Pool into an equal concentration



Adapter1 + Barcode + Read1 + Gene primer



Adapter2 + Read2 + Gene primer

Read1

Read2

Potential sources of bias in the library preparation:

- DNA extraction
- Gene primer
- Labeled primers
- PCR conditions
- PCR clean-up protocol

Advantages of this approach

- Culture-independent
- Many samples processed simultaneously
- Phylogenetically informative (databases exist)

Flow chart of 16 S amplicon analysis

Raw Fastq reads

↓

Quality trimming

↓

Merge R1 & R2 reads (if sequenced paired-end)

↓

Clustering reads into OTUs

↓

Organize into an OTU table

↓

Higher-level analyses (e.g. phylo tree)

Amplicon processing – what steps to take?

Issue	Notes	
Controls (positive and negative)	Negative: extraction and/ or PCR blanks	Positive: mock community, but what kind?
Pair reads or not?	One direction: which one? Quality issues	Bias in which reads were paired?
OTU binning	Against a reference database or de novo?	Which algorithm to use?
Taxonomy assignments	Which method?	What quality assignment to accept?
Low abundance OTUs	Are those real?	What is the cut-off?

Amplicon processing – what steps to take?

Issue	Notes	
Controls (positive and negative)	Negative: extraction and/ or PCR blanks	Positive: mock community, but what kind?
Pair reads or not?	One direction: which one? Quality issues	Bias in which reads were paired?
OTU binning	Against a reference database or de novo?	Which algorithm to use?
Taxonomy assignments	Which method?	What quality assignment to accept?
Low abundance OTUs	Are those real?	What is the cut-off?

Once you've decided on the steps, have to decide how to implement them

Analysis Tools

QIIME: http://qiime.org/tutorials/index.html

http://oligotyping.org

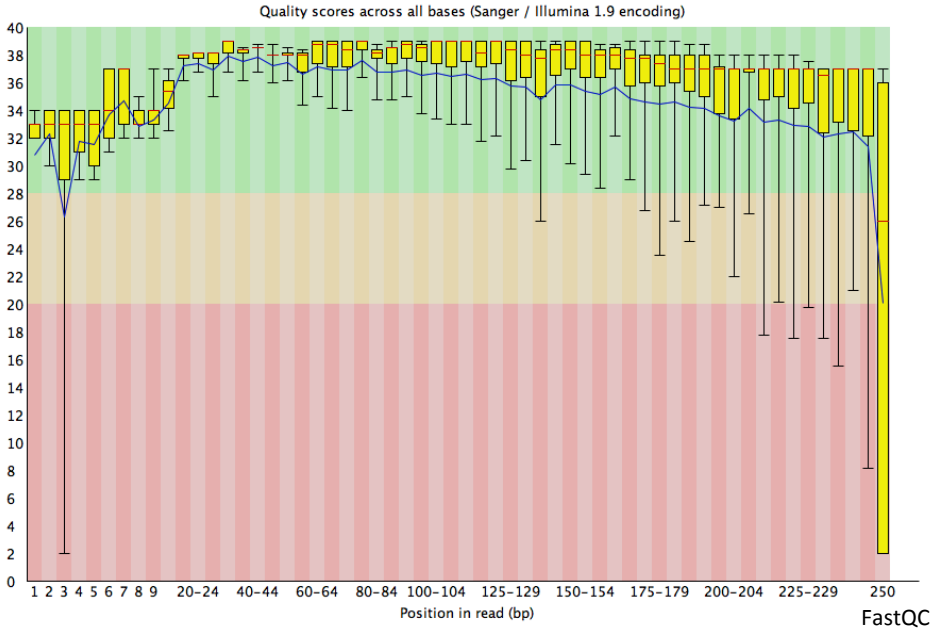
mothur: mothur.org

FastX-Toolkit: http://hannonlab.cshl.edu/fastx_toolkit/

UPARSE: http://drive5.com/uparse/

R: http://www.r-project.org

Quality of R1 and R2: R1 looks pretty good



FastQC

Quality of R1 and R2: R2 looks pretty bad