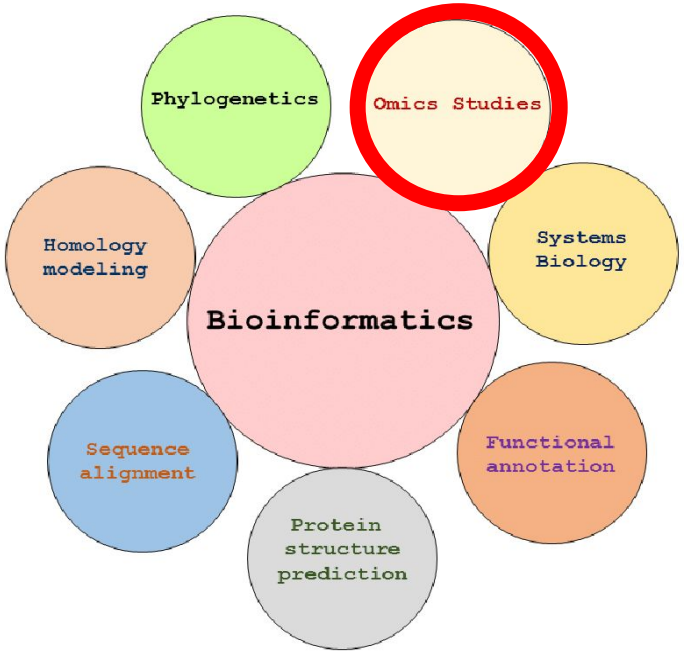

Metagenomics
and
amplicon sequencing

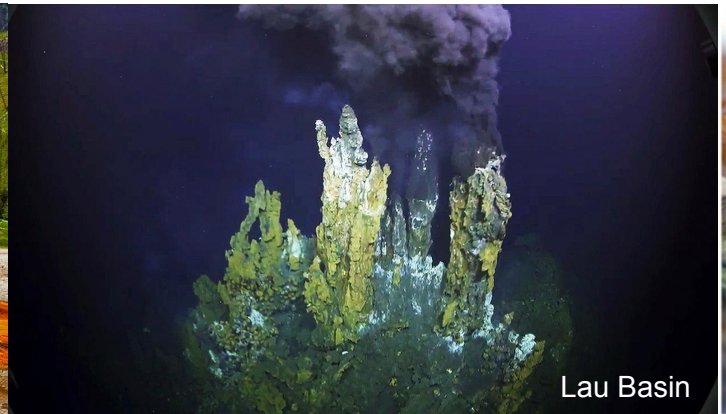
What is bioinformatics?

- Combines biology and computer science to **interpret biological data**



Field trip!

1. Pick a field destination/ system of your choice
2. Sample for some bacteria

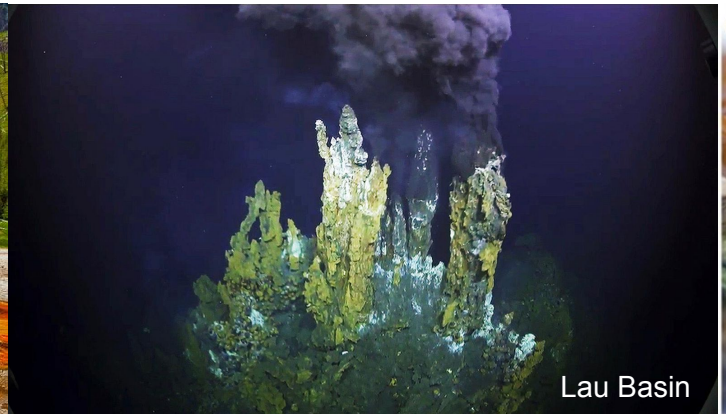


Field trip!

1. Pick a field destination/ system of your choice
2. Sample for some bacteria

Questions you may ask yourself:

Who is there? And what could they be doing?!



WHO is there?

WHAT could they be doing?

WHO is there?

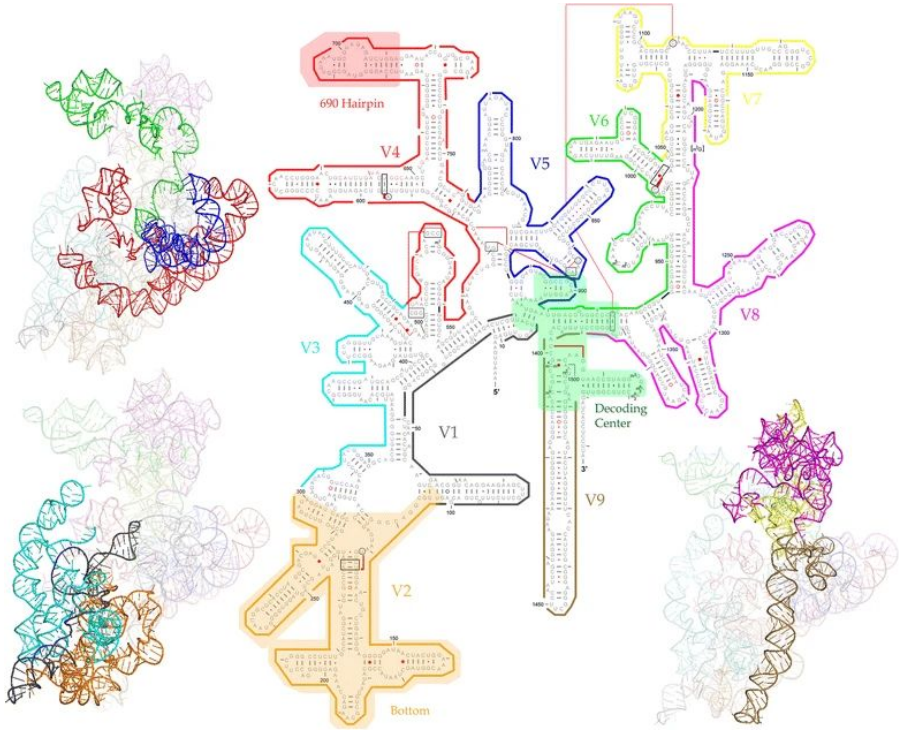
- We often attack this question with **amplicon sequencing**

WHAT could they be doing?

Amplicon

- Working with an amplified gene target
- Usually using SSU rRNA
  - 16S rRNA - prokaryotes
  - 18S rRNA - eukaryotes
  - Very conserved target among different domains of life
- Benefits:
  - Highly targeted
  - Faster
  - Less expensive
- Drawbacks:
  - Not looking at whole genomes

16S rRNA



Yang et al., 2016

Linear map of the 9 hypervariable regions of 16S rRNA

- Primers target specific regions
- Can sequence whole sequence (1500bp) - Sanger sequencing
    - Downsides: more computationally expensive
- Next generation sequencing sequences small chunks/regions
- 16S rRNA prefered regions: V4, V5, V6

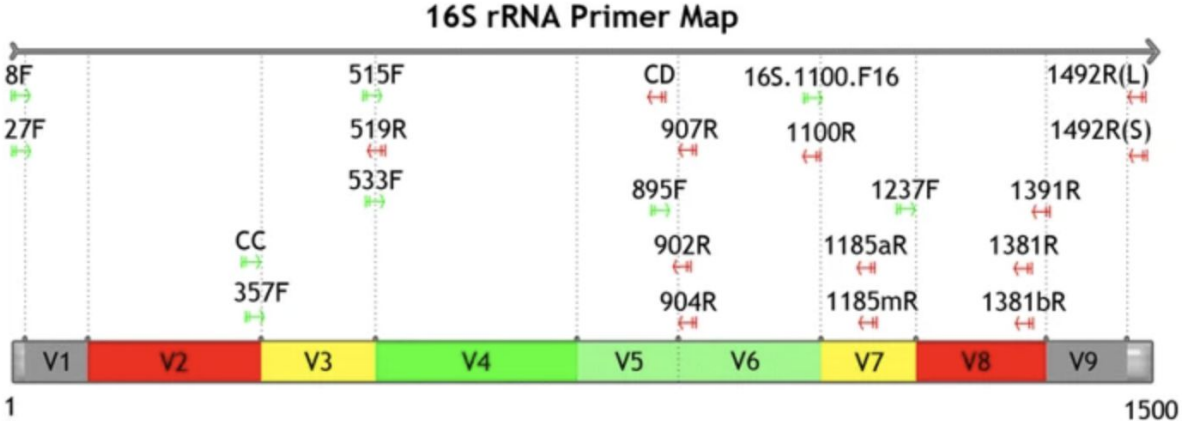
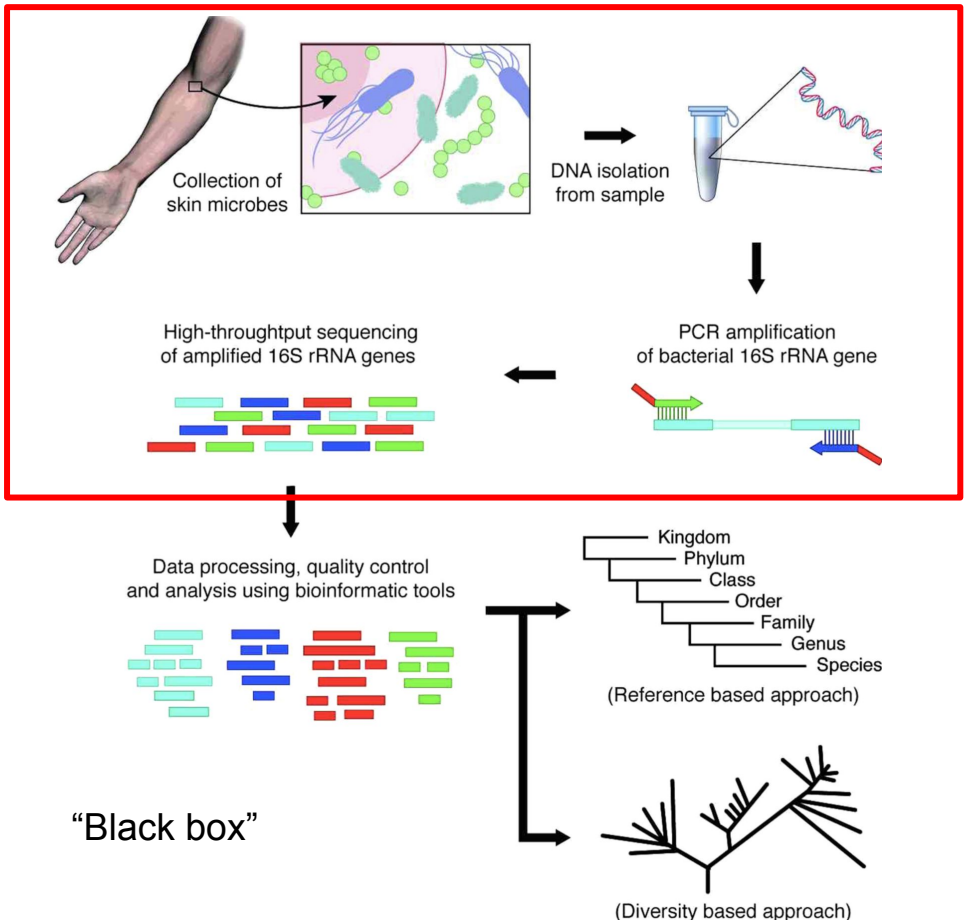


Illustration of different variable regions. Red regions (V2, V8) have a poor phylogenetic resolution at the phylum level. Green regions (V4, V5, V6) are associated with the shortest geodesic distance, which suggests that they may be the best choice for phylogeny-related analyses and the phylogenetic analysis of novel bacterial phyla. The figure refers to the primer map from Lutzonilab (http://lutzonilab.org/16s-ribosomal-dna/). Use of this information was approved by the original authors of the website

Yang et al., 2016

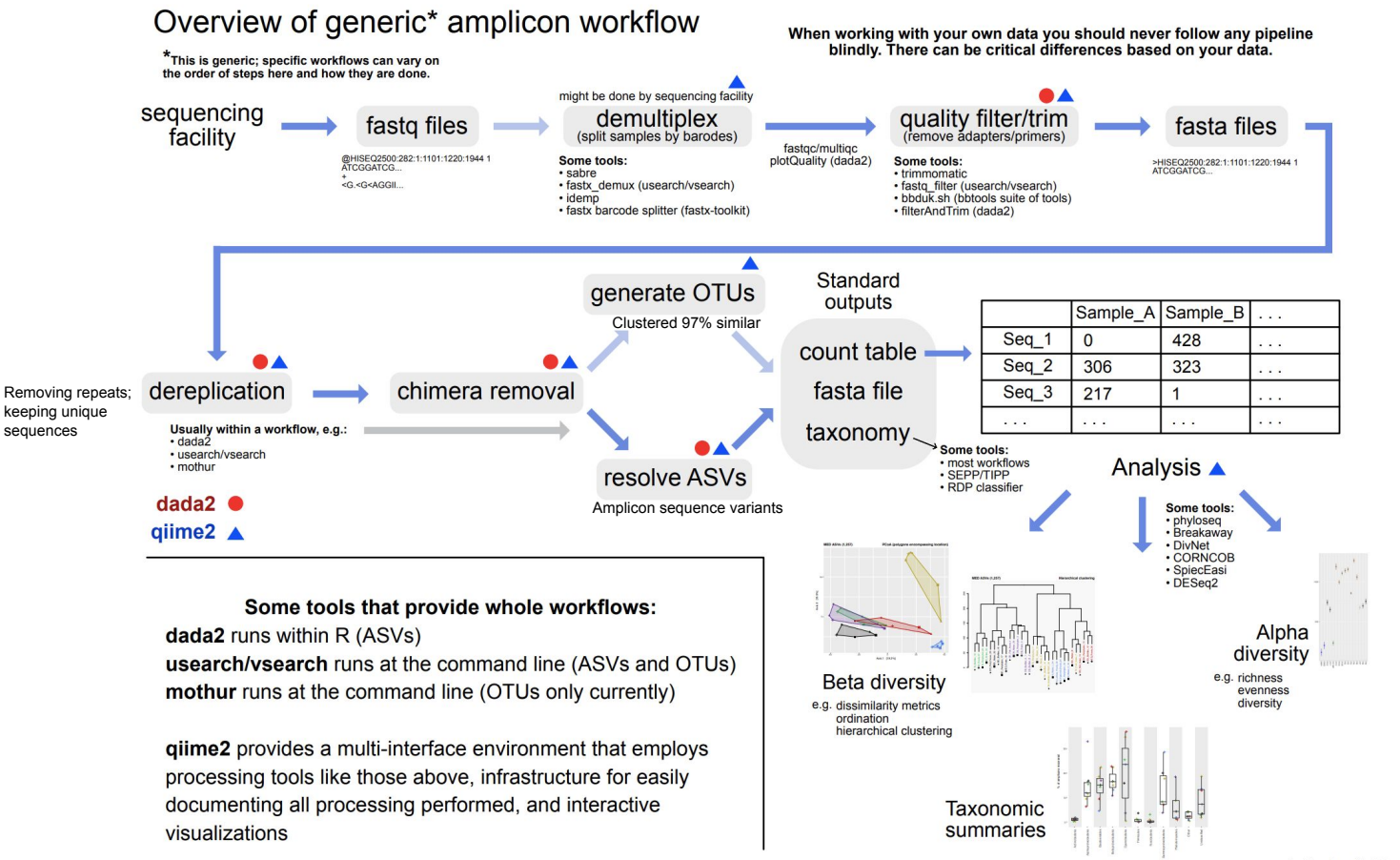
Amplicon sequencing: Lab-based workflow



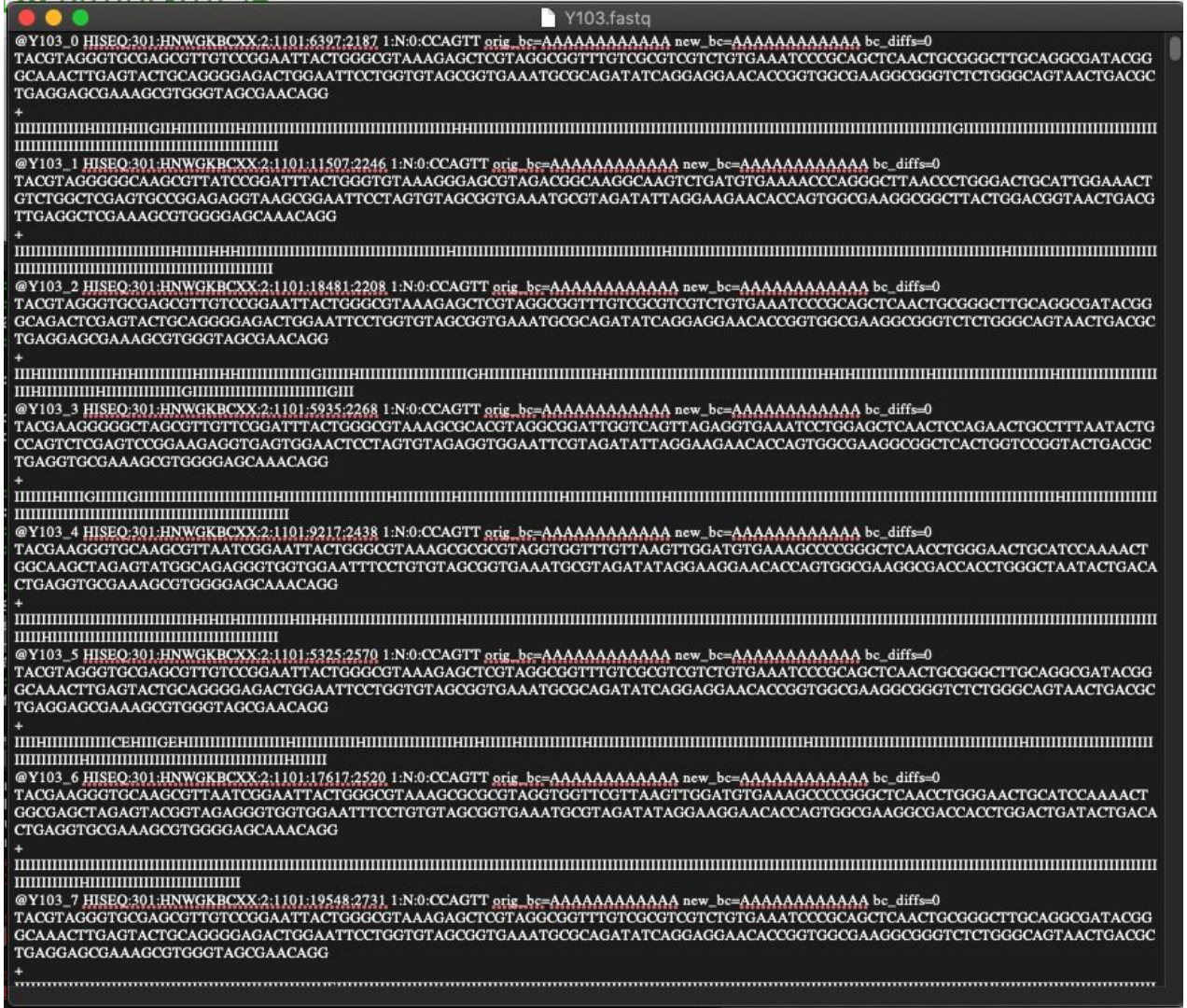
"Black box"

Jo et al., 2016

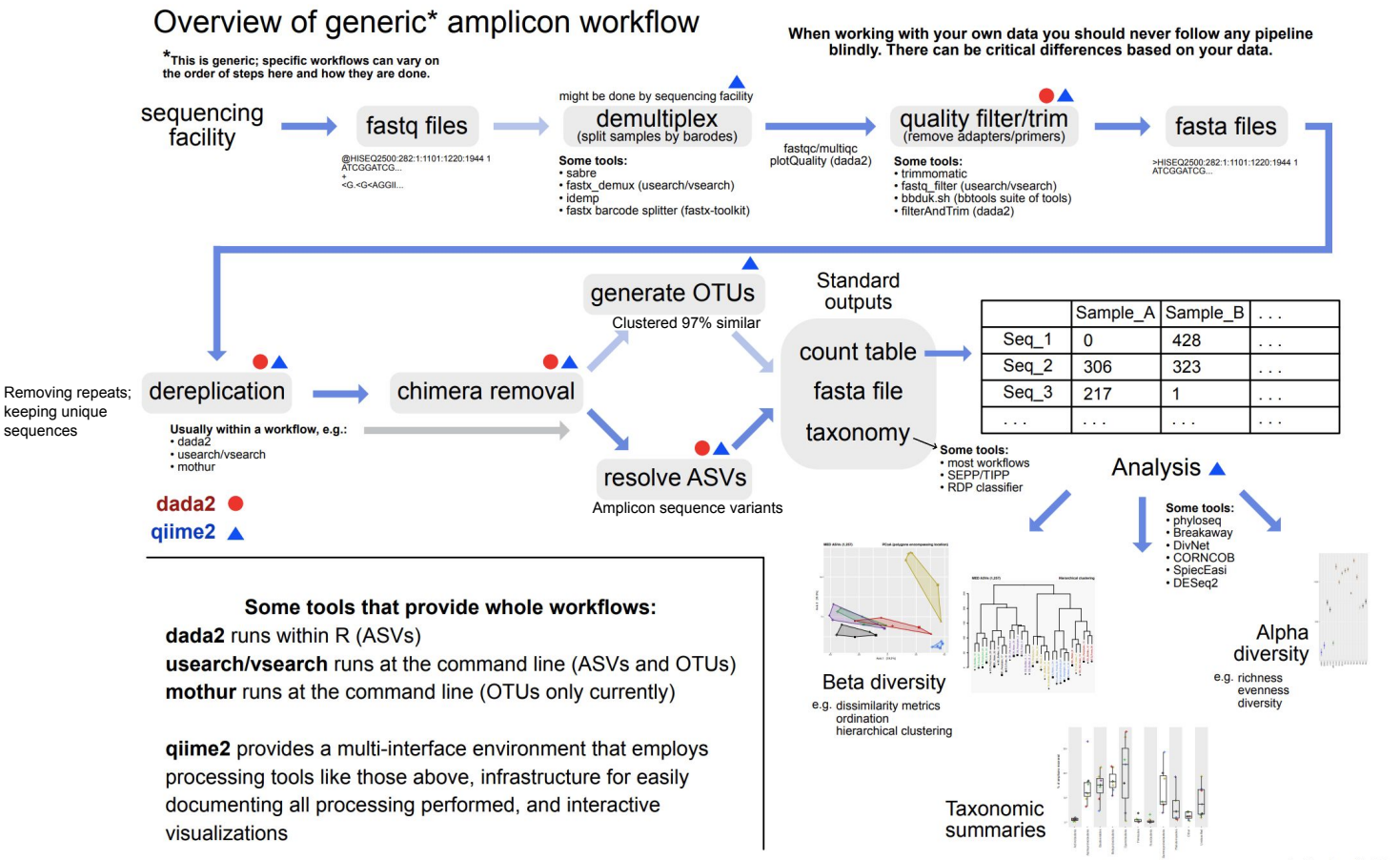
Amplicon sequencing: Bioinformatic workflow



astrobiomike.github.io


Y103.fastq
@Y103_0 HISEQ:301:HNWGKBCXX:2:1101:6397:2187 1:N:0:CCAGTT orig_bc=AAAAAAAAAAAA new_bc=AAAAAAAAAAAA bc_diffs=0
TACGTAGGGTGCGAGCGTTGTCCGGAATTACTGGGCGTAAAGAGCTCGTAGGCGGTTTGTCGCGTCGTCTGTGAAATCCCGCAGCTCAACTGCGGGCTTGCAGGCGATACGG
GCAAACTTGAGTACTGCAGGGGAGACTGGAATTCCTGGTGTAGCGGTGAAATGCGCAGATATCAGGAGGAACACCGGTGGCGAAGGCGGGTCTCTGGGCAGTAACTGACGC
TGAGGAGCGAAAGCGTGGGTAGCGAACAGG
+
@Y103_1 HISEQ:301:HNWGKBCXX:2:1101:11507:2246 1:N:0:CCAGTT orig_bc=AAAAAAAAAAAA new_bc=AAAAAAAAAAAA bc_diffs=0
TACGTAGGGGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGCAAGGCAAGTCTGATGTGAAAACCCAGGGCTTAACCCTGGGACTGCATTGGAAACT
GTCTGGCTCGAGTGCCGGAGAGGTAAGCGGAATTCCTAGTGTAGCGGTGAAATGCGTAGATATTAGGAAGAACACCAGTGGCGAAGGCGGCTTACTGGACGGTAACTGACG
TTGAGGCTCGAAAGCGTGGGGAGCAAACAGG
+
@Y103_2 HISEQ:301:HNWGKBCXX:2:1101:18481:2208 1:N:0:CCAGTT orig_bc=AAAAAAAAAAAA new_bc=AAAAAAAAAAAA bc_diffs=0
TACGTAGGGTGCGAGCGTTGTCCGGAATTACTGGGCGTAAAGAGCTCGTAGGCGGTTTGTCGCGTCGTCTGTGAAATCCCGCAGCTCAACTGCGGGCTTGCAGGCGATACGG
GCAGACTCGAGTACTGCAGGGGAGACTGGAATTCCTGGTGTAGCGGTGAAATGCGCAGATATCAGGAGGAACACCGGTGGCGAAGGCGGGTCTCTGGGCAGTAACTGACGC
TGAGGAGCGAAAGCGTGGGTAGCGAACAGG
+
@Y103_3 HISEQ:301:HNWGKBCXX:2:1101:5935:2268 1:N:0:CCAGTT orig_bc=AAAAAAAAAAAA new_bc=AAAAAAAAAAAA bc_diffs=0
TACGAAGGGGGCTAGCGTTGTTCGGATTTACTGGGCGTAAAGCGCACGTAGGCGGATTGGTCAGTTAGAGGTGAAATCCTGGAGCTCAACTCCAGAACTGCCTTTAATACTG
CCAGTCTCGAGTCCGGAAGAGGTGAGTGGAACTCCTAGTGTAGAGGTGGAATTCGTAGATATTAGGAAGAACACCAGTGGCGAAGGCGGCTCACTGGTCCGGTACTGACGC
TGAGGTGCGAAAGCGTGGGGAGCAAACAGG
+
@Y103_4 HISEQ:301:HNWGKBCXX:2:1101:9217:2438 1:N:0:CCAGTT orig_bc=AAAAAAAAAAAA new_bc=AAAAAAAAAAAA bc_diffs=0
TACGAAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCGCGTAGGTGGTTTGTTAAGTTGGATGTGAAAGCCCCGGGCTCAACCTGGGAACTGCATCCAAAACT
GGCAAGCTAGAGTATGGCAGAGGGTGGTGGAATTTCCTGTGTAGCGGTGAAATGCGTAGATATAGGAAGGAACACCAGTGGCGAAGGCGACCACCTGGGCTAATACTGACA
CTGAGGTGCGAAAGCGTGGGGAGCAAACAGG
+
@Y103_5 HISEQ:301:HNWGKBCXX:2:1101:5325:2570 1:N:0:CCAGTT orig_bc=AAAAAAAAAAAA new_bc=AAAAAAAAAAAA bc_diffs=0
TACGTAGGGTGCGAGCGTTGTCCGGAATTACTGGGCGTAAAGAGCTCGTAGGCGGTTTGTCGCGTCGTCTGTGAAATCCCGCAGCTCAACTGCGGGCTTGCAGGCGATACGG
GCAAACTTGAGTACTGCAGGGGAGACTGGAATTCCTGGTGTAGCGGTGAAATGCGCAGATATCAGGAGGAACACCGGTGGCGAAGGCGGGTCTCTGGGCAGTAACTGACGC
TGAGGAGCGAAAGCGTGGGTAGCGAACAGG
+
@Y103_6 HISEQ:301:HNWGKBCXX:2:1101:17617:2520 1:N:0:CCAGTT orig_bc=AAAAAAAAAAAA new_bc=AAAAAAAAAAAA bc_diffs=0
TACGAAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCGCGTAGGTGGTTCGTTAAGTTGGATGTGAAAGCCCCGGGCTCAACCTGGGAACTGCATCCAAAACT
GGCGAGCTAGAGTACGGTAGAGGGTGGTGGAATTTCCTGTGTAGCGGTGAAATGCGTAGATATAGGAAGGAACACCAGTGGCGAAGGCGACCACCTGGACTGATACTGACA
CTGAGGTGCGAAAGCGTGGGGAGCAAACAGG
+
@Y103_7 HISEQ:301:HNWGKBCXX:2:1101:19548:2731 1:N:0:CCAGTT orig_bc=AAAAAAAAAAAA new_bc=AAAAAAAAAAAA bc_diffs=0
TACGTAGGGTGCGAGCGTTGTCCGGAATTACTGGGCGTAAAGAGCTCGTAGGCGGTTTGTCGCGTCGTCTGTGAAATCCCGCAGCTCAACTGCGGGCTTGCAGGCGATACGG
GCAAACTTGAGTACTGCAGGGGAGACTGGAATTCCTGGTGTAGCGGTGAAATGCGCAGATATCAGGAGGAACACCGGTGGCGAAGGCGGGTCTCTGGGCAGTAACTGACGC
TGAGGAGCGAAAGCGTGGGTAGCGAACAGG
+

Amplicon sequencing: Bioinformatic workflow



astrobiomike.github.io

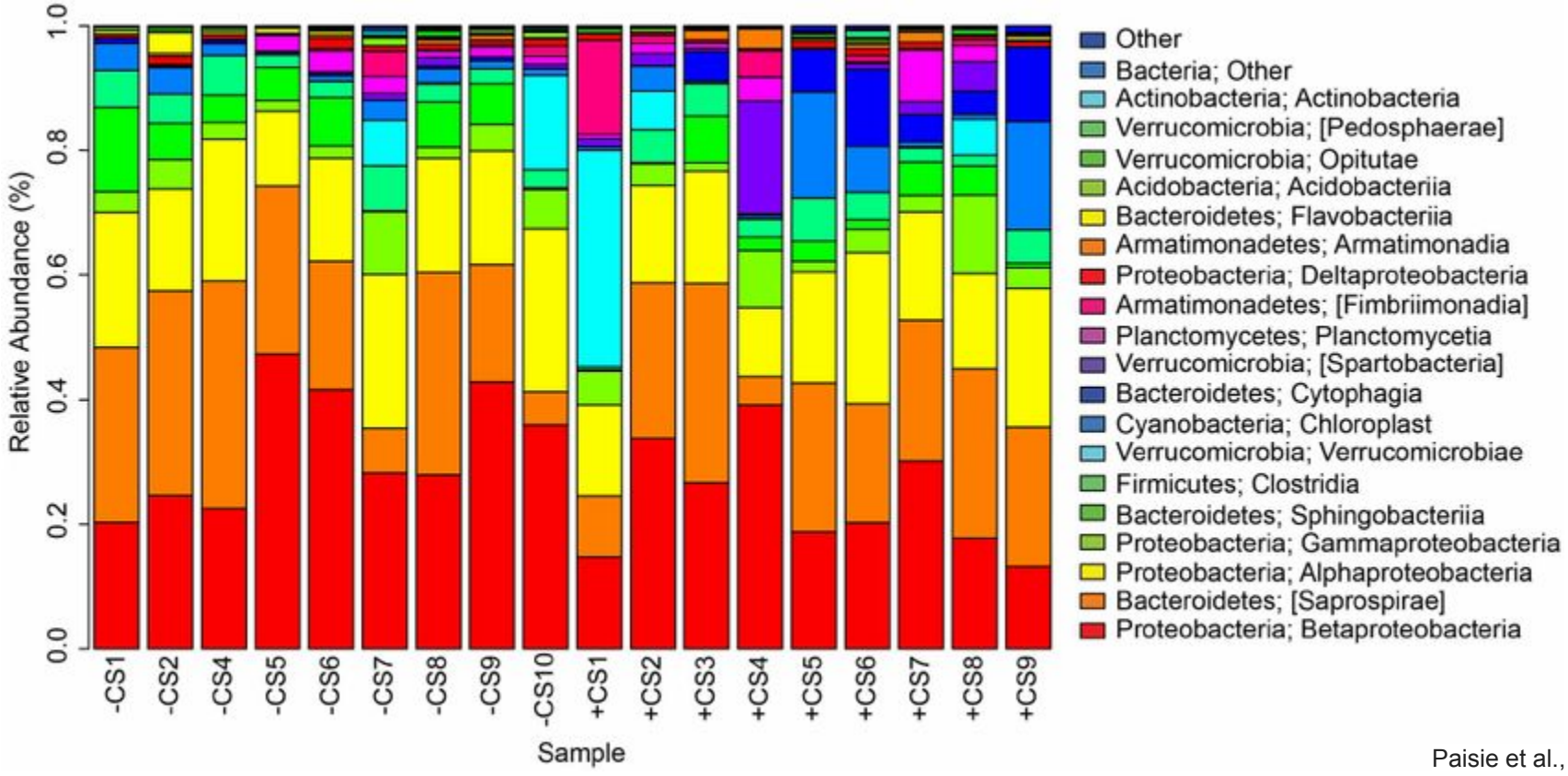
Reference databases

It's important to know and understand what reference database you are using, especially for taxonomy assignment.

Database	Target Amplicon
SILVA	16S & 18S rRNA
RDP	16S rRNA & 28S rRNA (fungi)
PR2	18S rRNA
Greengenes	16S rRNA (*no longer maintained*)
UNITE	Fungal ITS

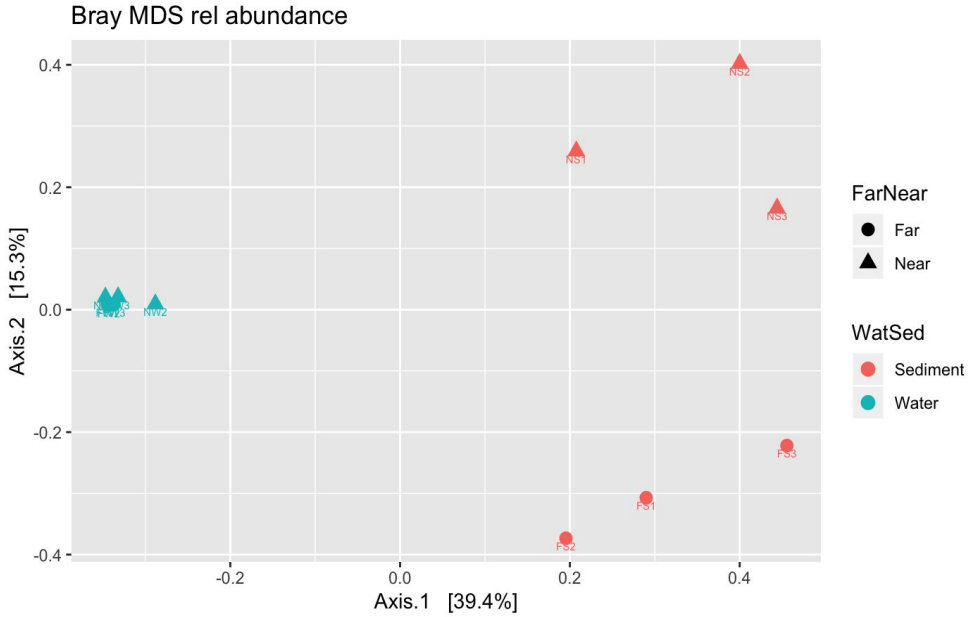
Slide credit: Liz Suter

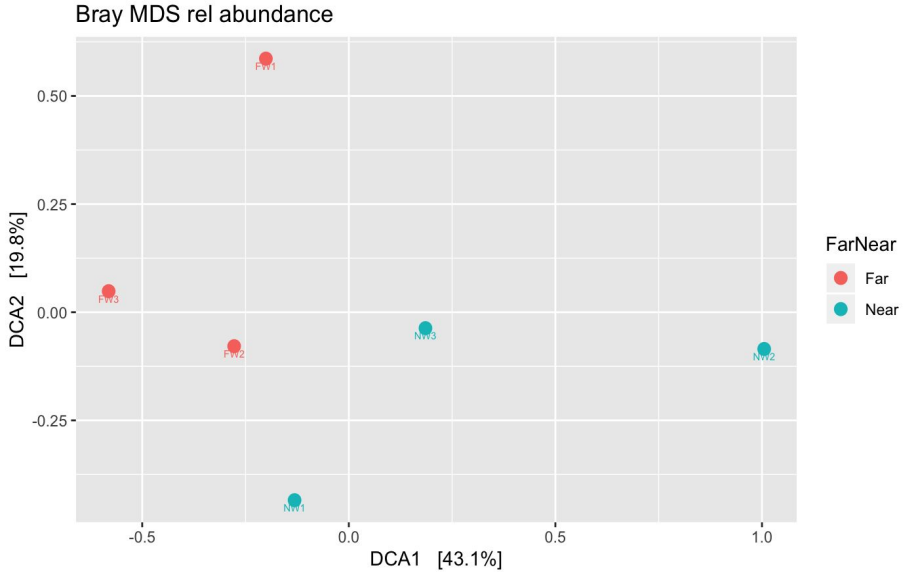
Example figures



Paisie et al., 2014

16S NMDS Plot





Questions?

WHO is there?

- We often attack this question with **amplicon sequencing**

WHAT could they be doing?

WHO is there?

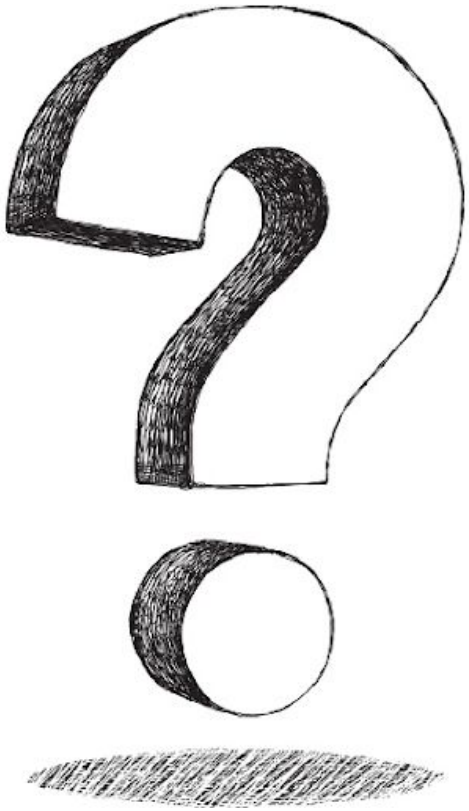
- We often attack this question with **amplicon sequencing**

WHAT could they be doing?

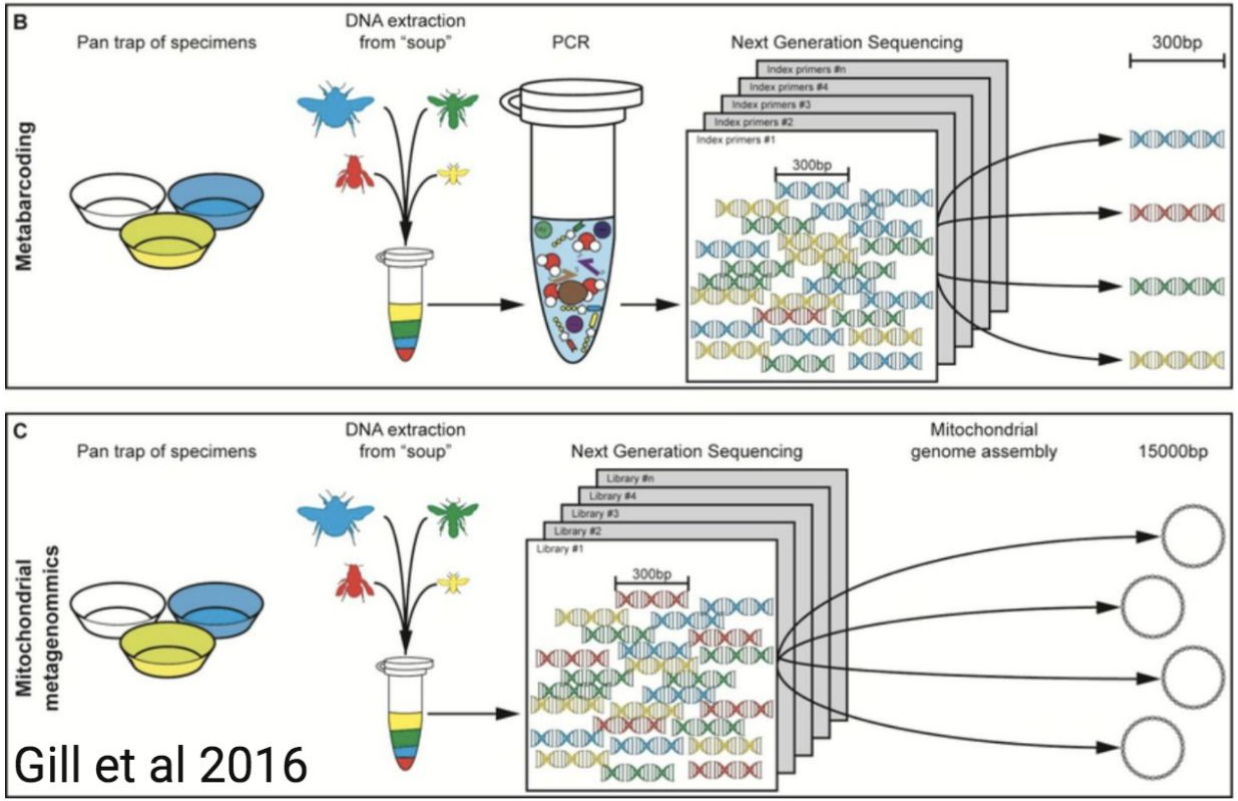
- We often attack this question with **metagenomic sequencing**

Metagenomic sequencing

- Also referred to as shotgun sequencing
- Looking at whole genomes
- Benefits:
  - LOTS of data
- Downside:
  - Expensive
  - Longer sequencing process



Amplicon sequencing vs. metagenomic “shotgun” sequencing

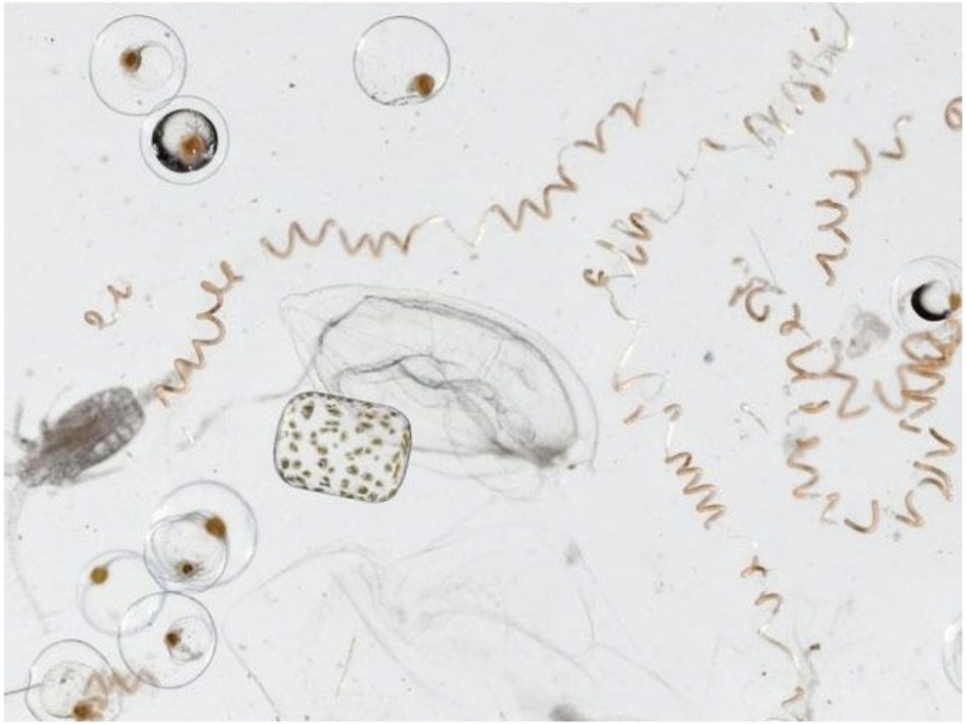


“Amplicon sequencing” or “metabarcoding” is sequencing a specific target region from many genomes (e.g. 16S rRNA gene, *nifH* gene)

“Shotgun Metagenomics” is (incomplete) sequencing of a mixture of genomes using an untargeted approach

Slide credit: Eric Collins

Incomplete because one drop of seawater contains about...



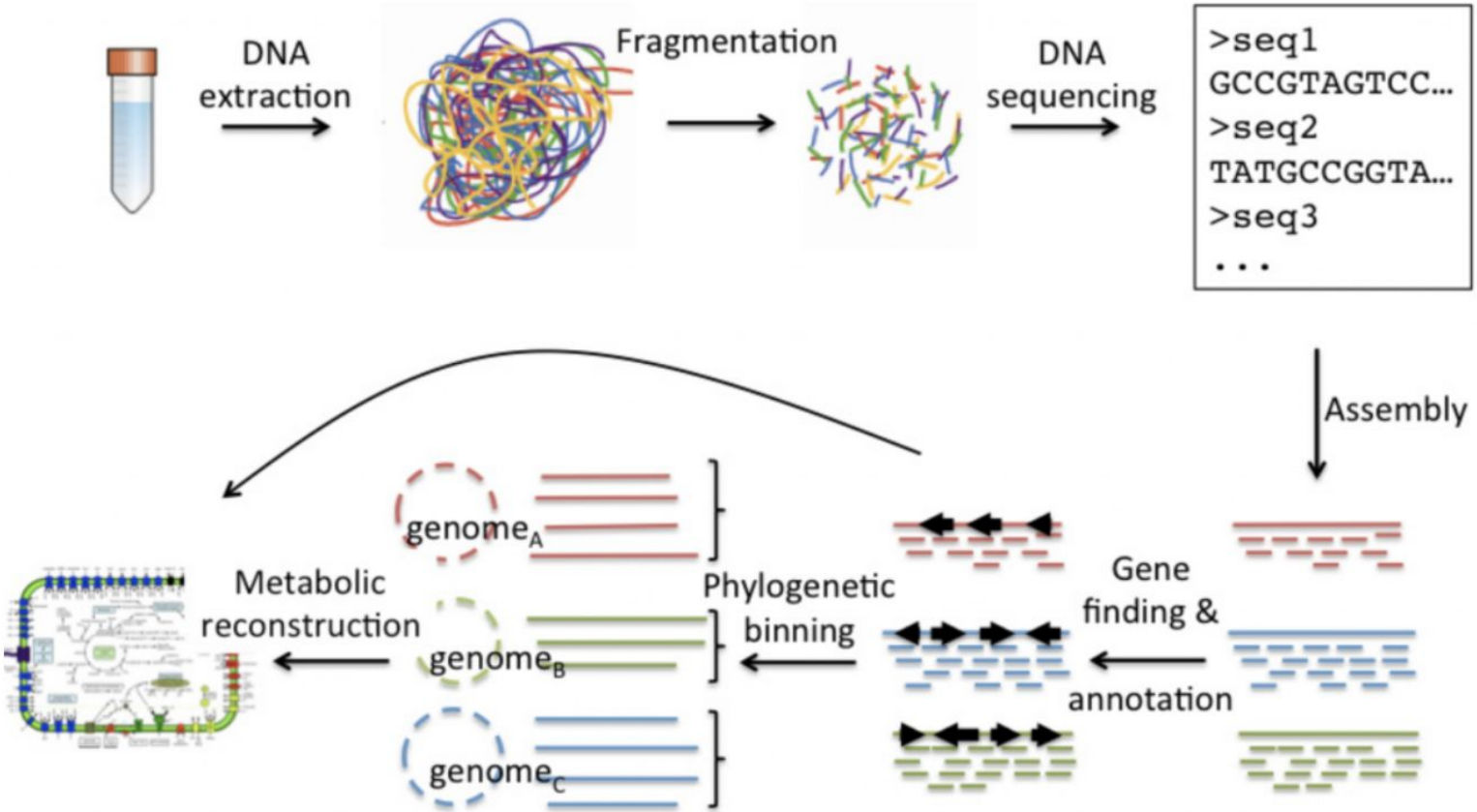
10^6 bacteria/mL * $3x10^6$ bp/bacteria = $3x10^{12}$ bp/mL

\+ 10^3 euks/mL * $3x10^8$ bp/euk = $3x10^{11}$ bp/mL

= 3.3 Tbp/mL

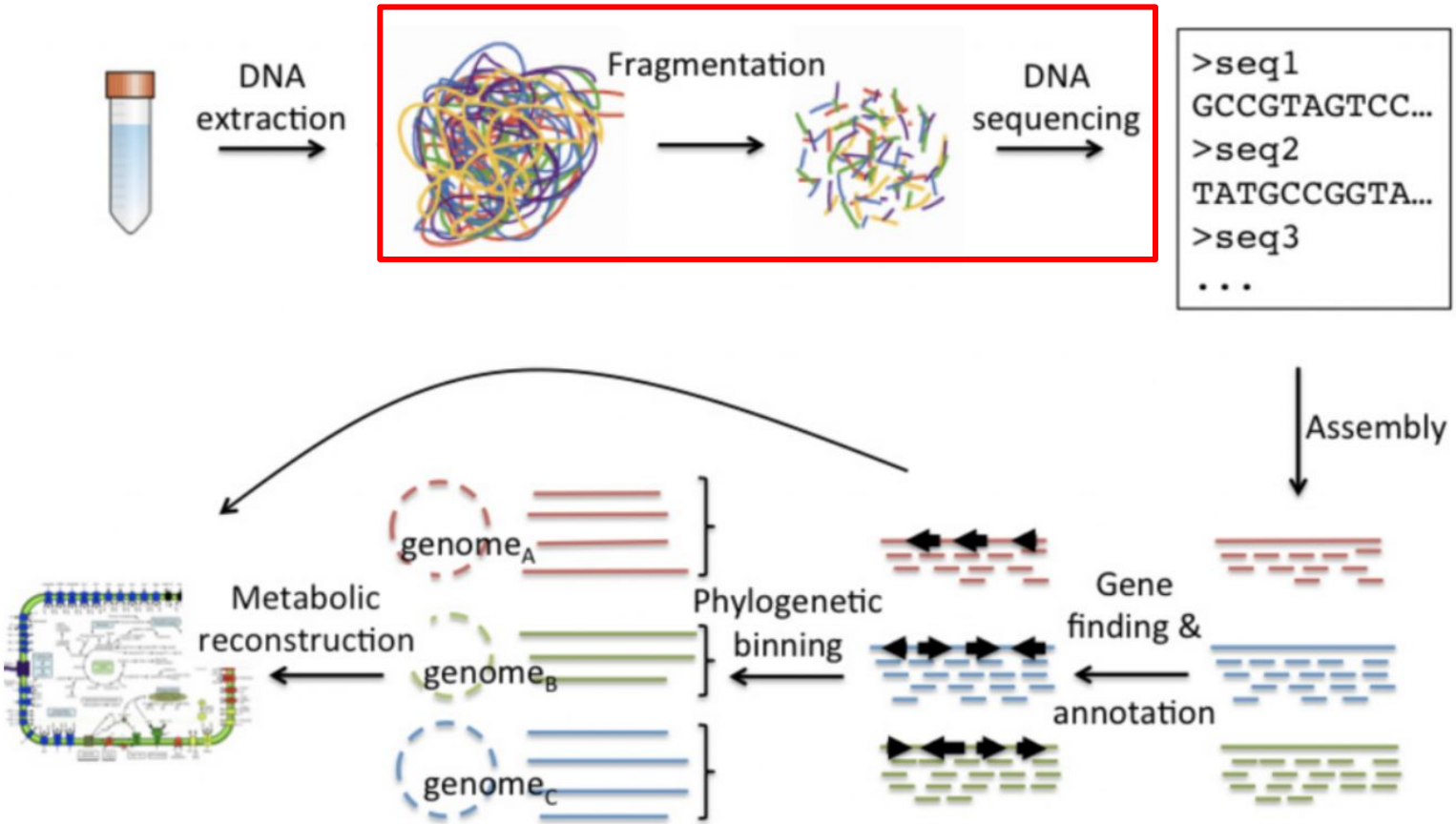
Slide credit: Eric Collins

Metagenomic workflow

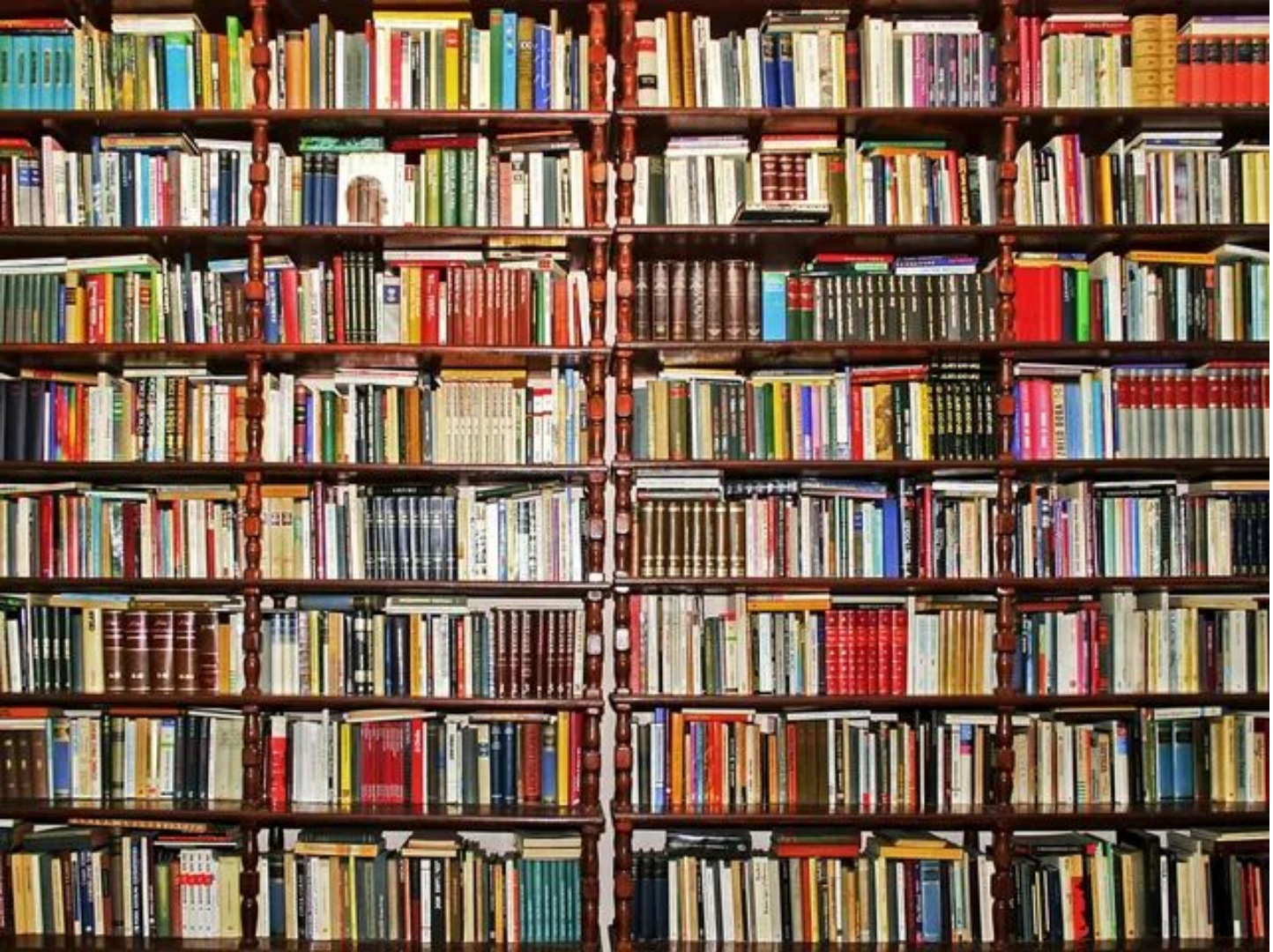


https://learn.gencore.bio.nyu.edu/metgenomics/

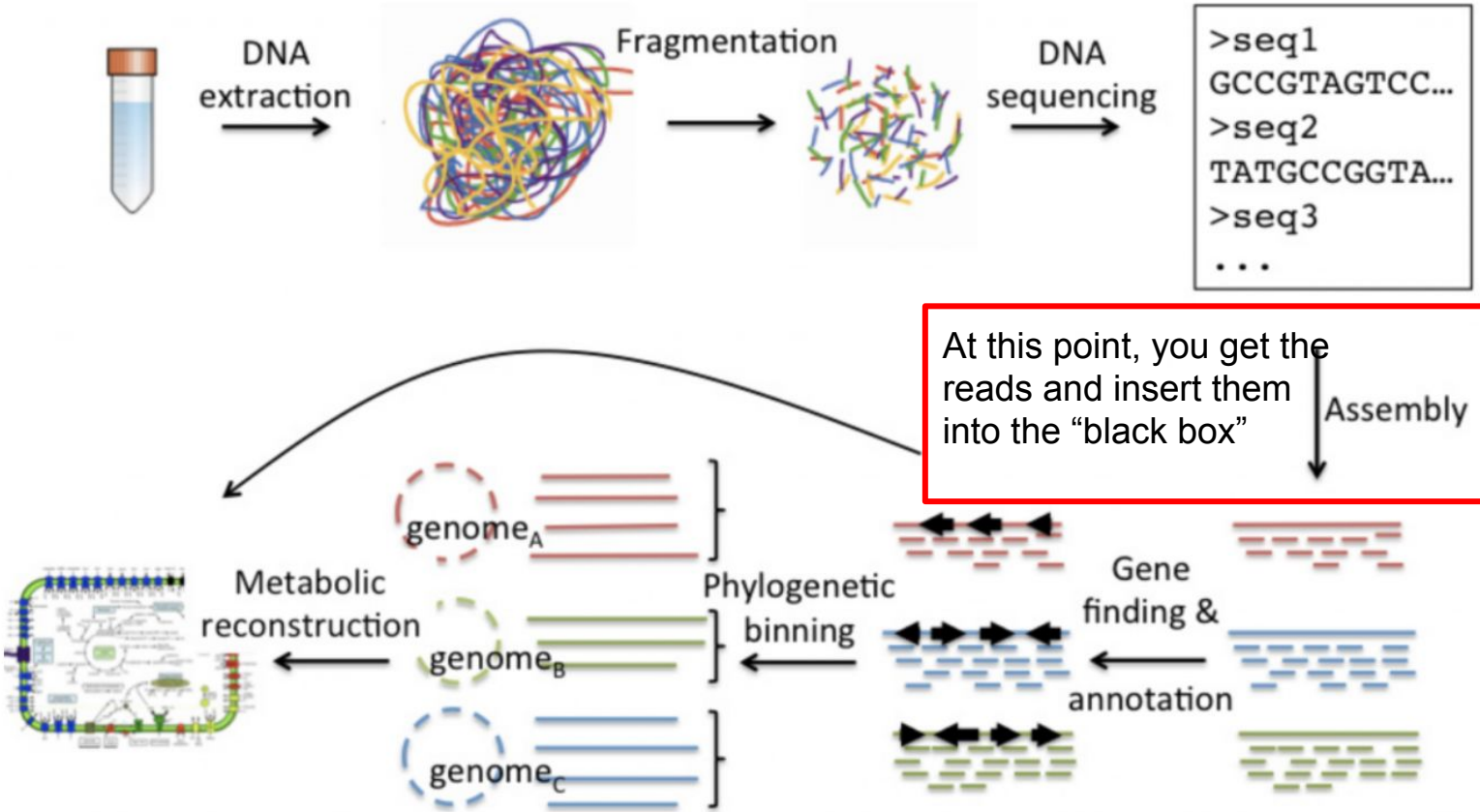
Metagenomic workflow



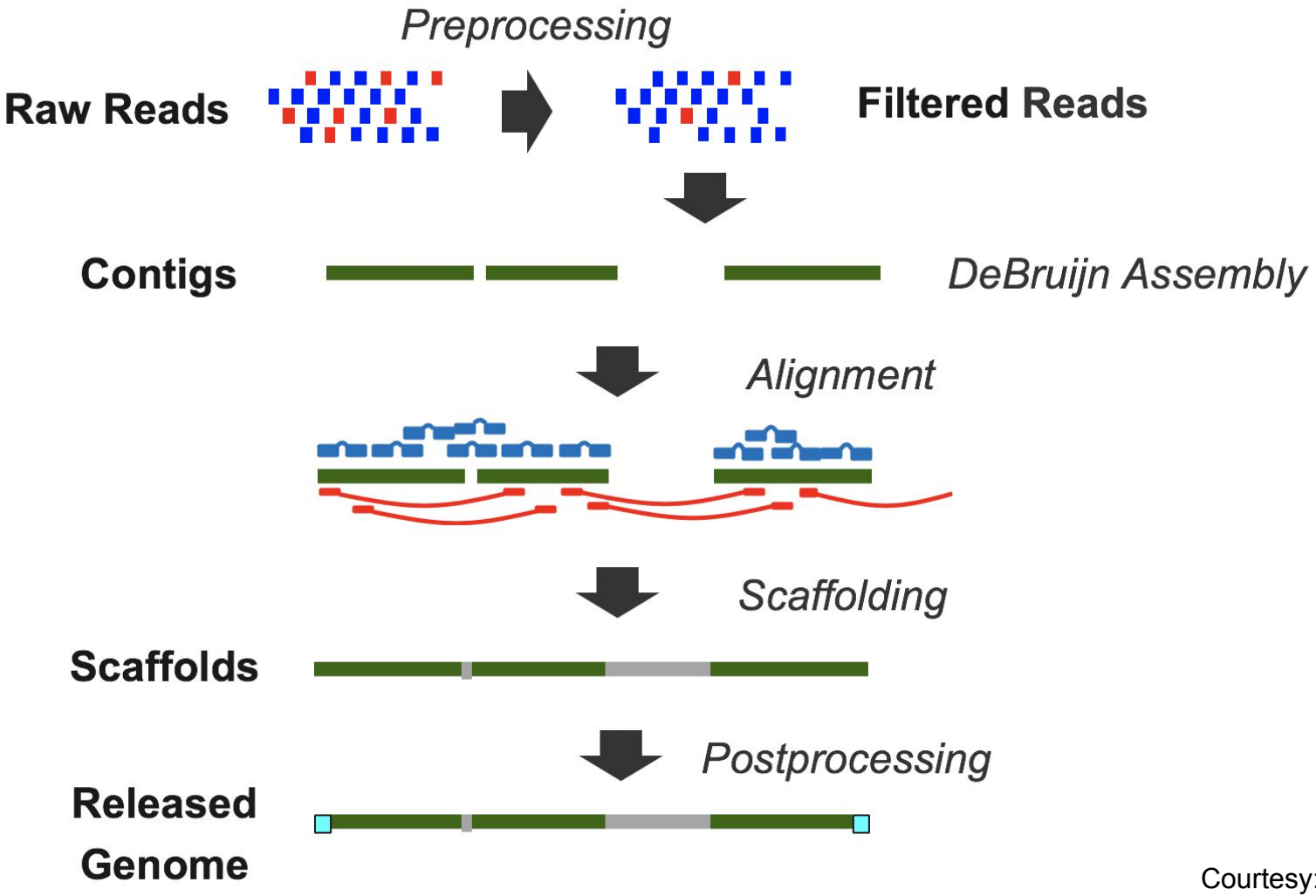
https://learn.gencore.bio.nyu.edu/metgenomics/



Metagenomic workflow



https://learn.gencore.bio.nyu.edu/metgenomics/



Courtesy: Brian Bushnell

“It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, ... “

Dickens, Charles. A Tale of Two Cities. 1859. London: Chapman Hall

Example courtesy of J. Leipzig 2010

itwasthebestoftimesitwastheworstoftimesitwastheageofwisdomitwastheageoffoolishness…

Generate random 'reads'   **How do we assemble?**

fincreduli geoffoolis ltwasthebe ltwasthebe geofwisdom itwastheep epochofinc timesitwas stheepocho nessitwast wastheageo theepochof stheepocho hofincredu estoftimes eoffoolish lishnessit hofbeliefi pochofincr itwasthewo twastheage toftimesit domitwasth ochofbelie eepochofbe eepochofbe astheworst chofincred theageofwi iefitwasth ssitwasthe astheepoch efitwasthe wisdomitwa ageoffooli twasthewor ochofbelie sdomitwast sitwasthea eepochofbe ffoolishne eofwisdomi hebestofti stheageoff twastheepo eworstofti stoftimesi theepochof esitwasthe heepochofi theepochof sdomitwast astheworst rstoftimes worstoftim stheepocho geoffoolis ffoolishne timesitwas lishnessit stheageoff eworstofti orstoftime fwisdomitw wastheageo heageofwis incredulit ishnessitw twastheepo wasthewors astheepoch heworstoft ofbeliefit wastheageo heepochofi pochofincr heageofwis stheageofw fincreduli astheageof wisdomitwa wastheageo astheepoch olishnessi astheepoch itwastheep twastheage wisdomitwa fbeliefitw bestoftime epochofbel theepochof sthebestof lishnessit hofbeliefi ltwasthebe ishnessitw sitwasthew ageofwisdo twastheage esitwasthe twastheage shnessitwa fincreduli fbeliefitw theepochof mesitwasth domitwasth ochofbelie heageofwis oftimesitw stheepocho bestoftime twastheage foolishnes ftimesitwa thebestoft itwastheag theepochof itwasthewo ofbeliefit bestoftime mitwasthea imesitwast timesitwas orstoftime estoftimes twasthebes stoftimesi sdomitwast wisdomitwa theworstof astheworst sitwasthew theageoffo eepochofbe

De Bruijn solution...

Courtesy: Brian Bushnell

Convert reads into “Kmers”

Kmer: a substring of defined length

…..etc for all reads in the dataset

Courtesy: Brian Bushnell

Using DNA sequences, it would look more like this...

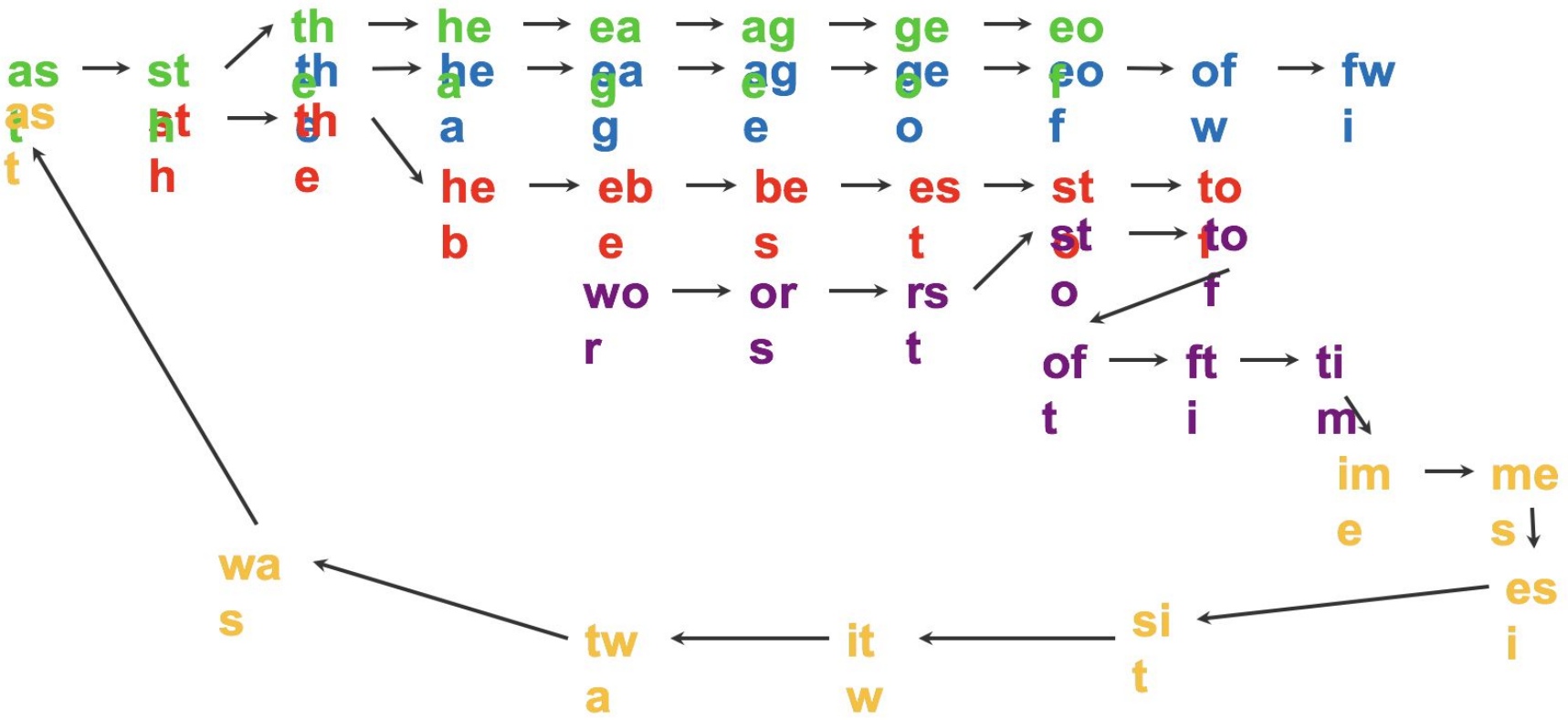
sequence **ATGGAAGTCGCGGAATC**

7mers

ATGGAAG
TGGAAGT
GGAAGTC
GAAGTCG
AAGTCGC
AGTCGCG
GTCGCGG
TCGCGGA
CGCGGAA
GCGGAAT
CGGAATC

Modified from:
https://homolog.us/Tutorials/book4/p2.1.html

Build a De-Bruijn graph from the kmers

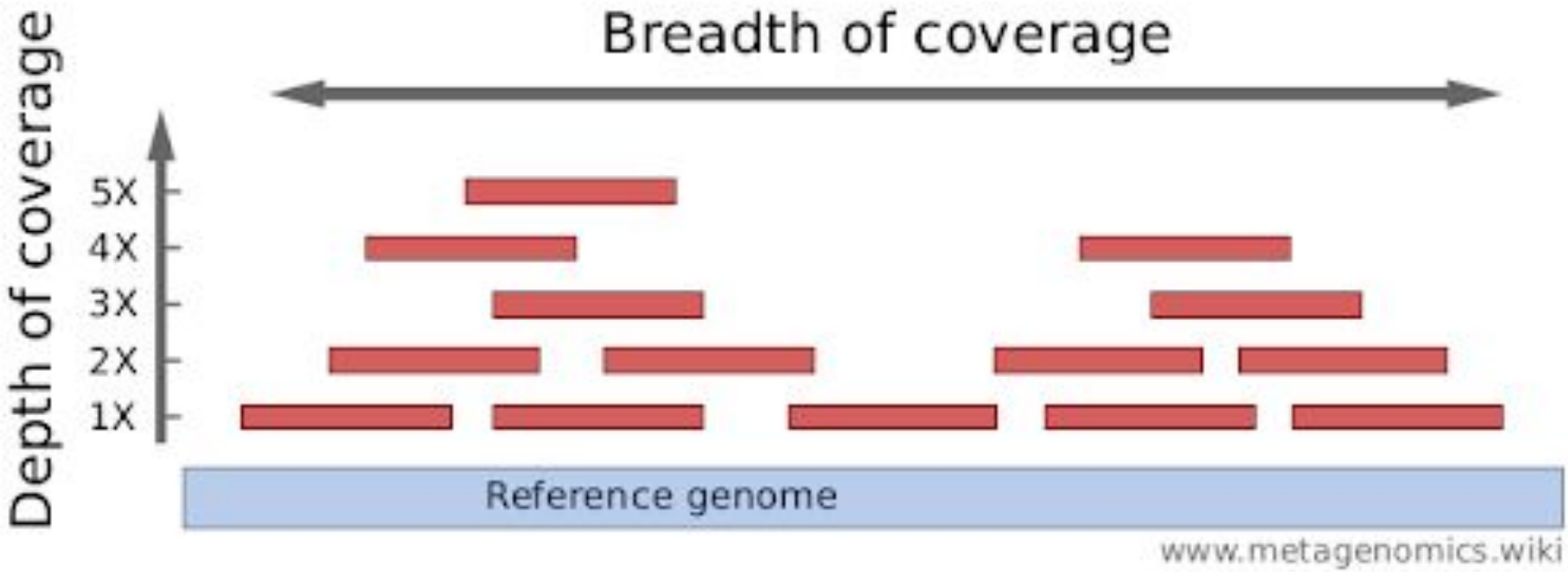


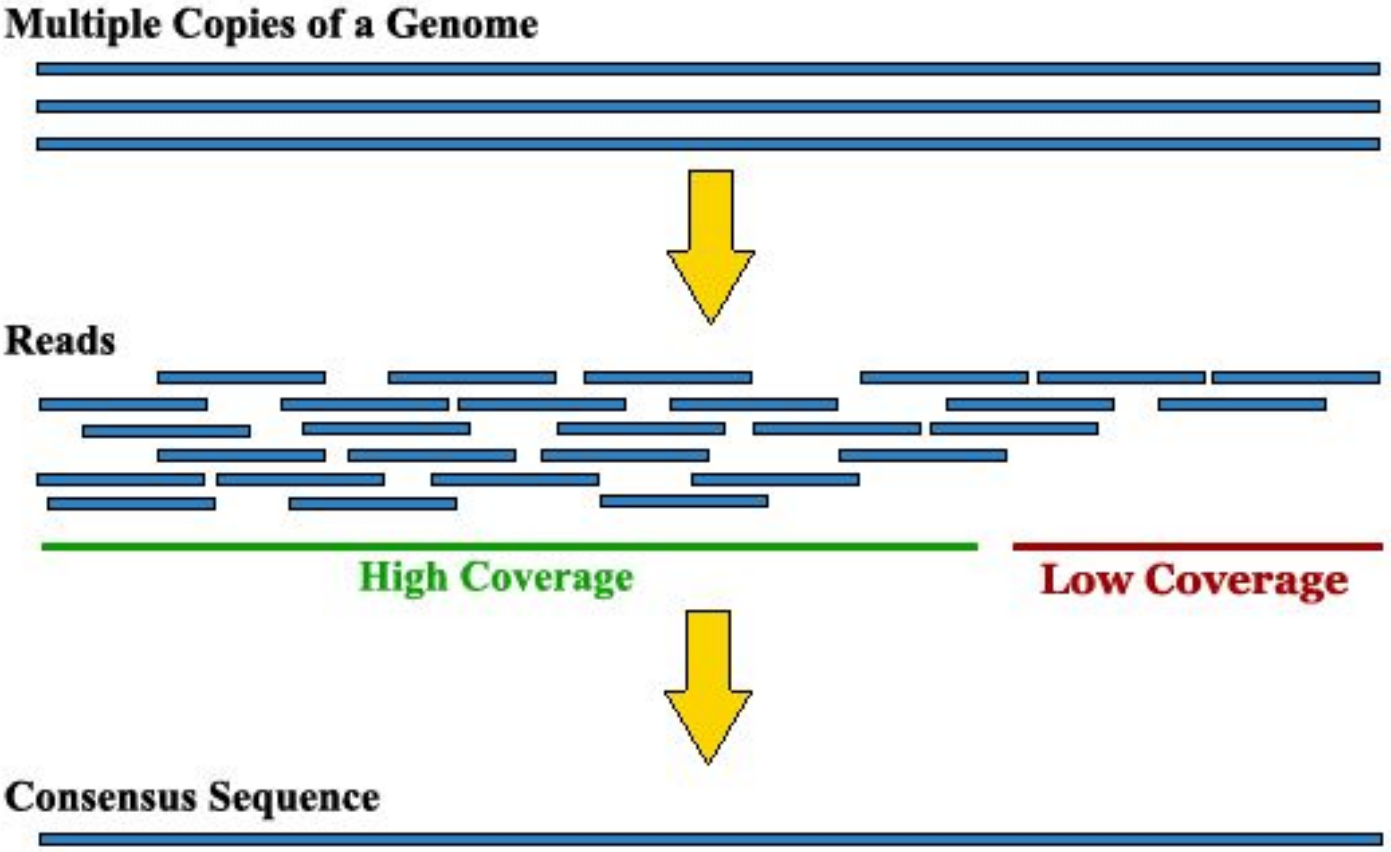
Courtesy: Brian Bushnell

“It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, ... “

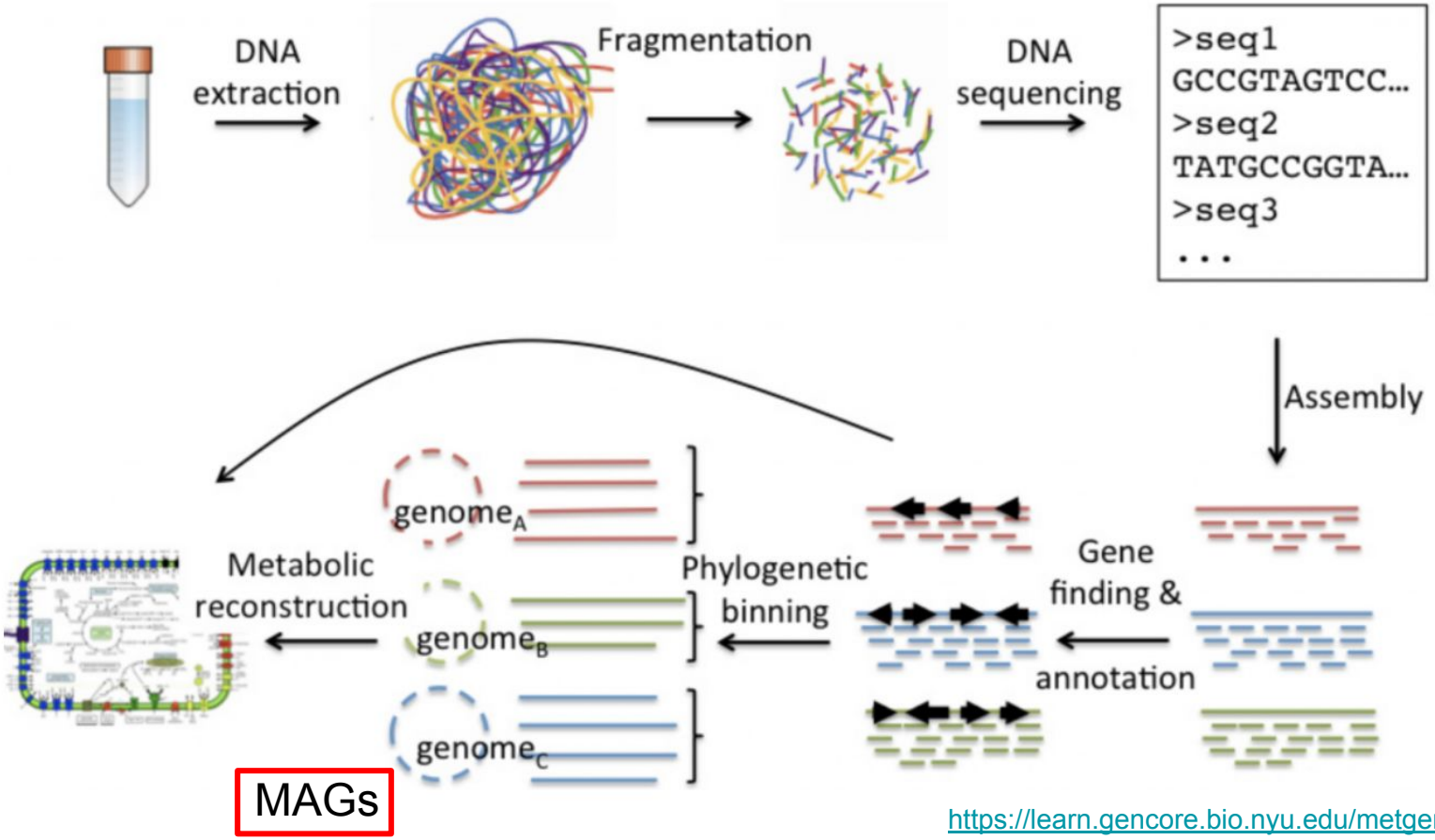
Dickens, Charles. A Tale of Two Cities. 1859. London: Chapman Hall

Example courtesy of J. Leipzig 2010


Breadth of coverage
Depth of coverage
5X
4X
3X
2X
1X
Reference genome
www.metagenomics.wiki


Multiple Copies of a Genome
Reads
High Coverage
Low Coverage
Consensus Sequence

Metagenomic workflow



https://learn.gencore.bio.nyu.edu/metgenomics/

Overview of generic metagenomics workflow

When working with your own data you should never follow any pipeline blindly. There can be critical differences based on your data.

sequencing facility → fastq files

@HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...
+
<G.<G<AGGII...

might be done by sequencing facility

demultiplex (split samples by barodes) → fastqc/multiqc → quality filter/trim (remove adapters/**primers**) → fasta files

>HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...

Some tools:
- sabre
- fastx_demux (usearch/vsearch)
- idemp
- fastx barcode splitter (fastx-toolkit)

Some tools:
- trimmomatic
- bbduk.sh (bbtools suite of tools)

read-based analysis

Some tools:
- metaphlan2
- humann2

no assembly path

assembly path

consider testing assemblies with and w/o

digital normalization

Some tools:
- bbnorm
- diginorm

MetaQUAST is a great tool for comparing assemblies

(co)-assembly

Some assemblers:
- SPAdes
- Megahit
- idba-ud

map individual sample reads to (co)-assembly

Generate coverage information (mappnig)

Some tools:
- bowtie2
- bwa

Gene calling
Functional/taxonomic profiling

Some tools:
- prodigal (identifies open reading frames)
- prokka (runs prodigal and performs annotations)
- GHOSTKOALA (web-hosted KEGG annotations)
- BLAST (protein nr db/refseq/COGs)

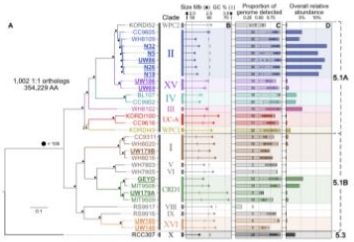
Recovering MAGs (metagenome-assembled genomes)

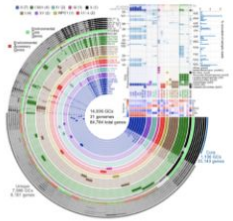
Some tools:
- anvi'o (interactive manual curation of bins)
- CONCOCT (kmer-based and coverage-based binning; also incorporated in anvi'o)
- MetaBAT2 (kmer-based and coverage-based)
- BinSanity (primarily coverage-based, optional second round kmer-based; memory intensive)
- checkm (taxonomy; estimate % completion/redundancy)

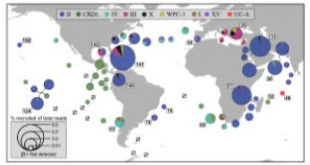
Some common genomics stuff

Phylogenomics
Comparative genomics
Pangenomics
Env. distributions

Some tools:
- anvi'o (integrated HMMs for common single-copy gene sets; integrated pangenomic workflow for identifying orthologs via OrthoMCL)
- PanOCT (identifies orthologs utilizing synteny information)
- MUSCLE (alignment software)
- FastTree (very fast, pseudo-maximum likelihood tree builder)
- RAxML (maximum likelihood tree builder)
- Mauve (whole-genome alignerment)







assemblies

(co)-assembly

map individual sample reads to (co)-assembly

Generate coverage information (mappnig)

Some assemblers:
- SPAdes
- Megahit
- idba-ud

Some tools:
- bowtie2
- bwa

Gene calling
Functional/taxonomic profiling

Recovering MAGs (metagenome-assembled genomes)

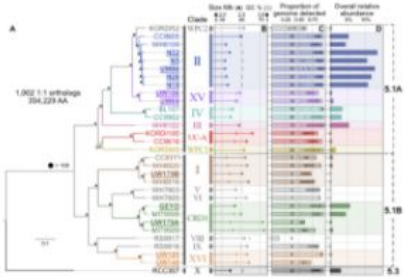
Some tools:
- prodigal (identifies open reading frames)
- prokka (runs prodigal and performs annotations)
- GHOSTKOALA (web-hosted KEGG annotations)
- BLAST (protein nr db/refseq/COGs)

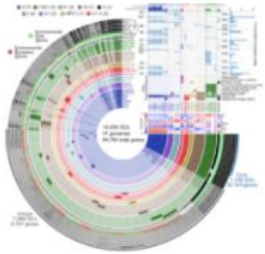
Some tools:
- anvi'o (interactive manual curation of bins)
- CONCOCT (kmer-based and coverage-based binning; also incorporated in anvi'o)
- MetaBAT2 (kmer-based and coverage-based)
- BinSanity (primarily coverage-based, optional second round kmer-based; memory intensive)
- checkm (taxonomy; estimate % completion/redundancy)

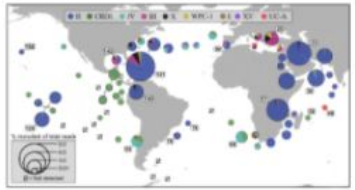
Some common genomics stuff

Phylogenomics
Comparative genomics
Pangenomics
Env. distributions

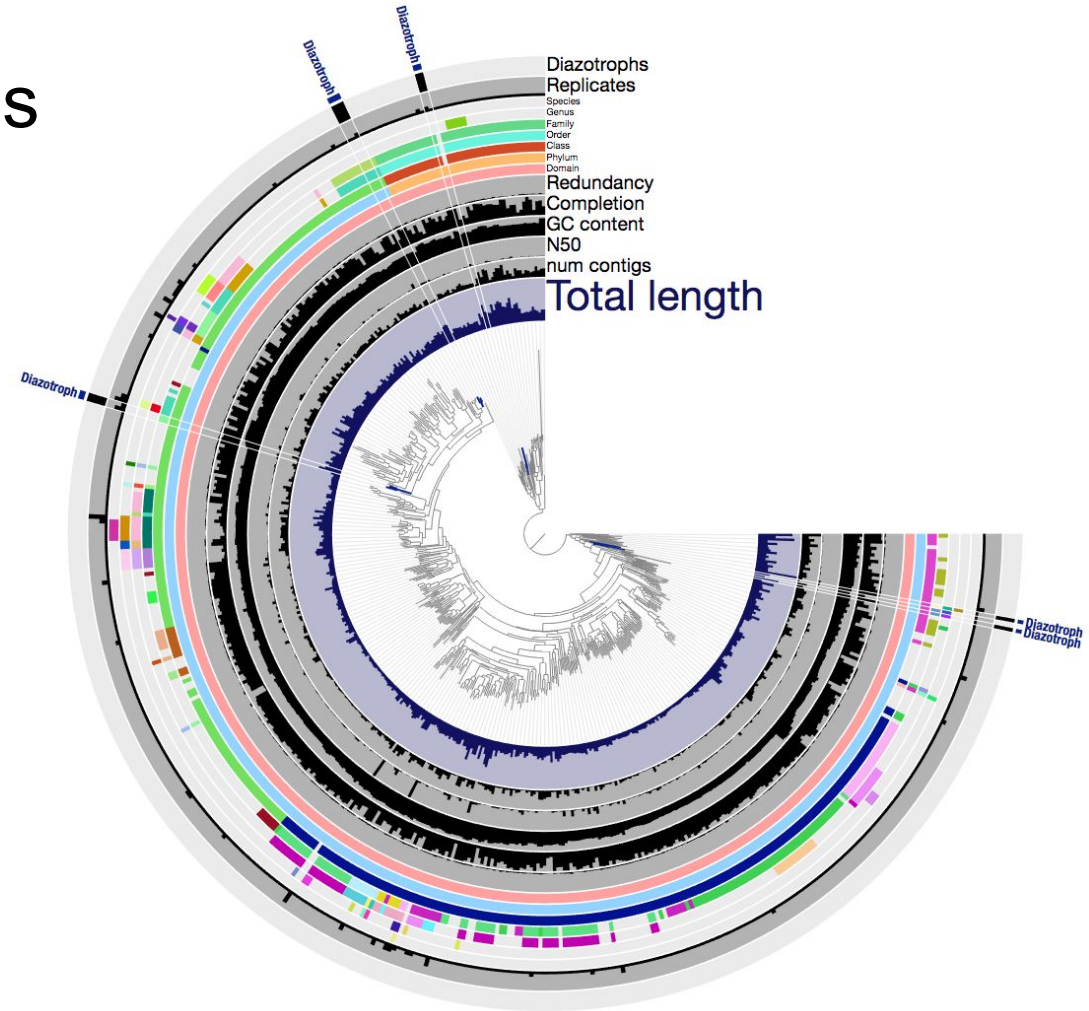
Some tools:
- anvi'o (integrated HMMs for common single-copy gene sets; integrated pangenomic workflow for identifying orthologs via OrthoMCL)
- PanOCT (identifies orthologs utilizing synteny information)
- MUSCLE (alignment software)
- FastTree (very fast, pseudo-maximum likelihood tree builder)
- RAxML (maximum likelihood tree builder)
- Mauve (whole-genome alignerment)

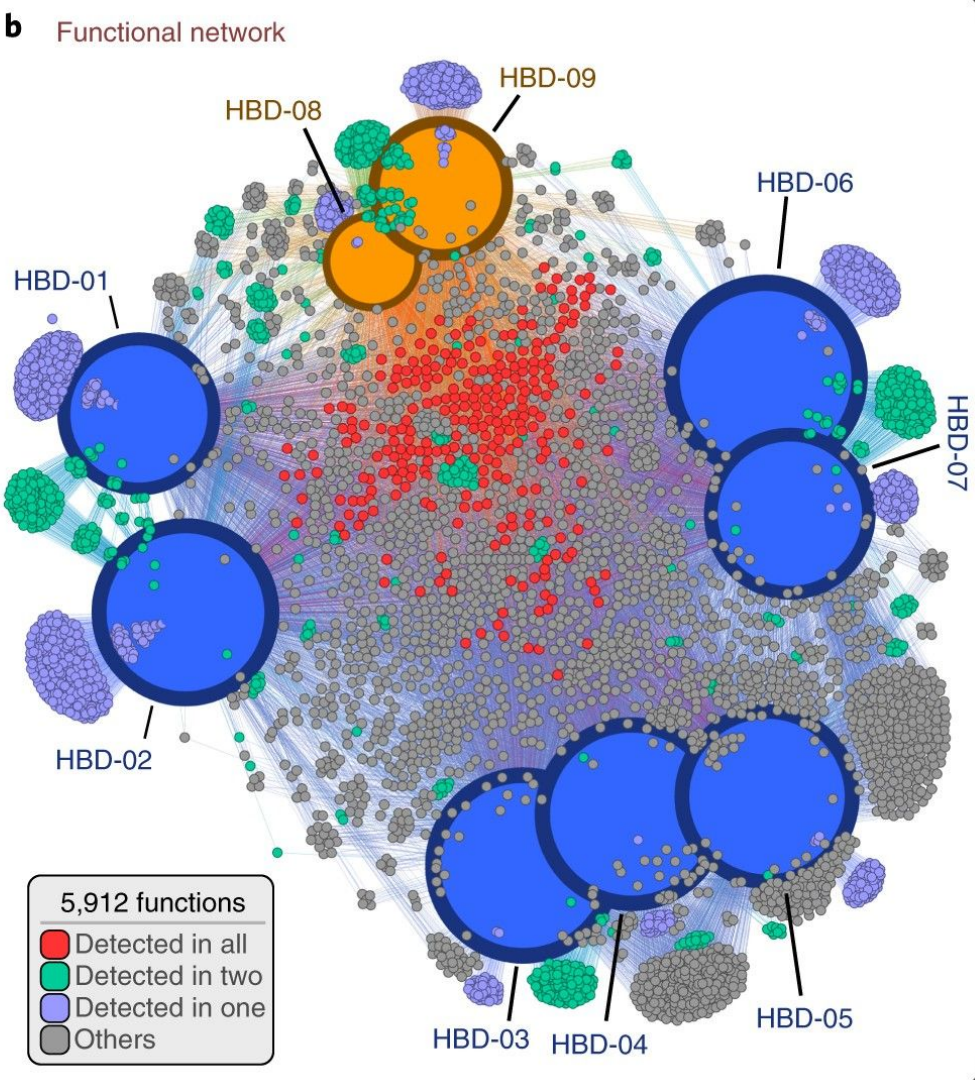






Example figures





Delmont et al., 2014

Recap: Metagenomic vs amplicon sequencing

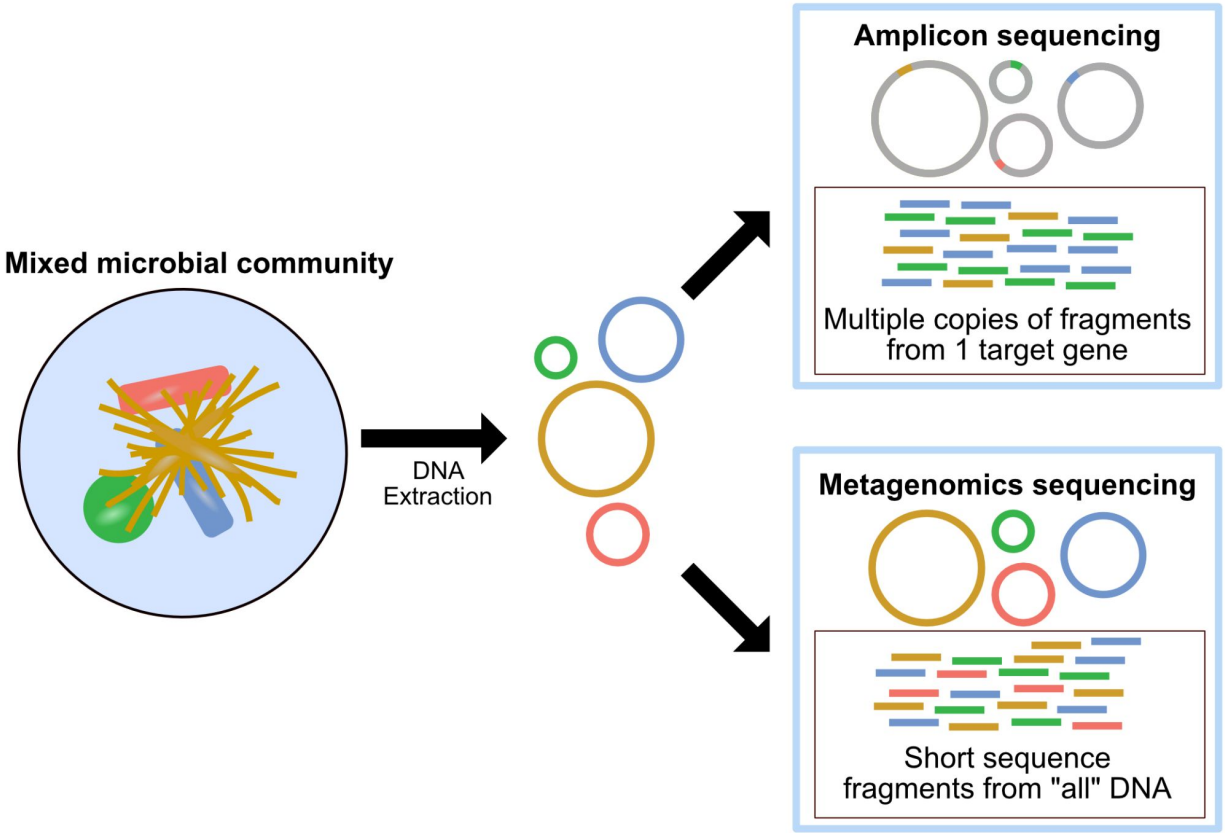


Image credit: Mike Lee

Metagenomic vs genomic sequencing

Metagenomic Sequencing

- **Purpose**: Metagenomic sequencing is used to analyze the **collective genome of a microbial community** directly from an environmental sample, without the need to culture organisms in the lab.

Genomic Sequencing

- **Purpose**: Genomic sequencing is focused on determining the complete DNA sequence of a **single organism's genome.**

Metagenomic vs genomic sequencing

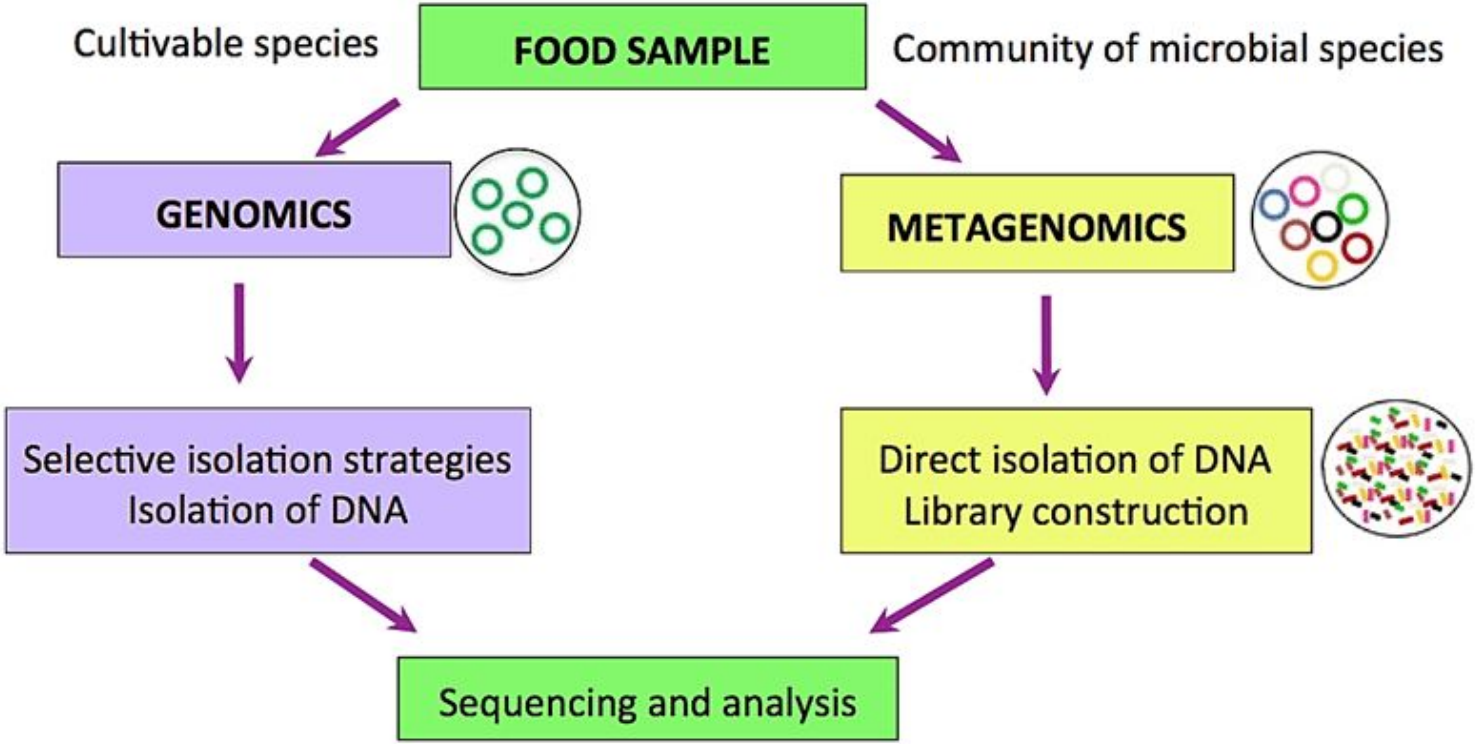
Metagenomic Sequencing

- **Purpose**: Metagenomic sequencing is used to analyze the **collective genome of a microbial community** directly from an environmental sample, without the need to culture organisms in the lab.
- Extract DNA from an entire sample with multiple community members

Genomic Sequencing

- **Purpose**: Genomic sequencing is focused on determining the complete DNA sequence of a **single organism's genome.**
- Usually requires isolating a single organism from a community of organisms and then extracting this DNA

Metagenomic vs genomic sequencing



Perrotta et al., 2017

Questions?