Metagenomics Taxonomic Classification & Assembly

Morgan Langille

Dalhousie University

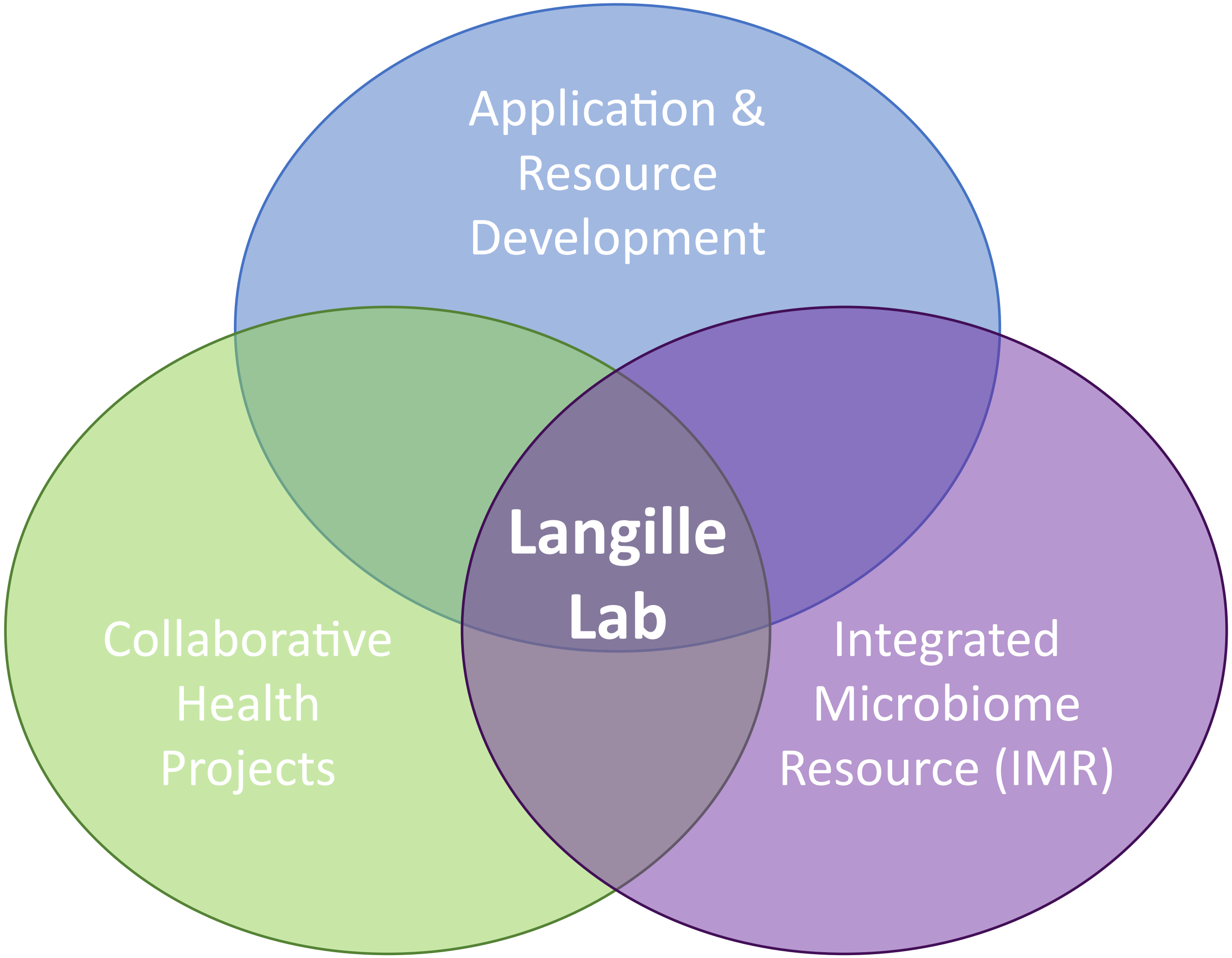
Dec. 7 2022

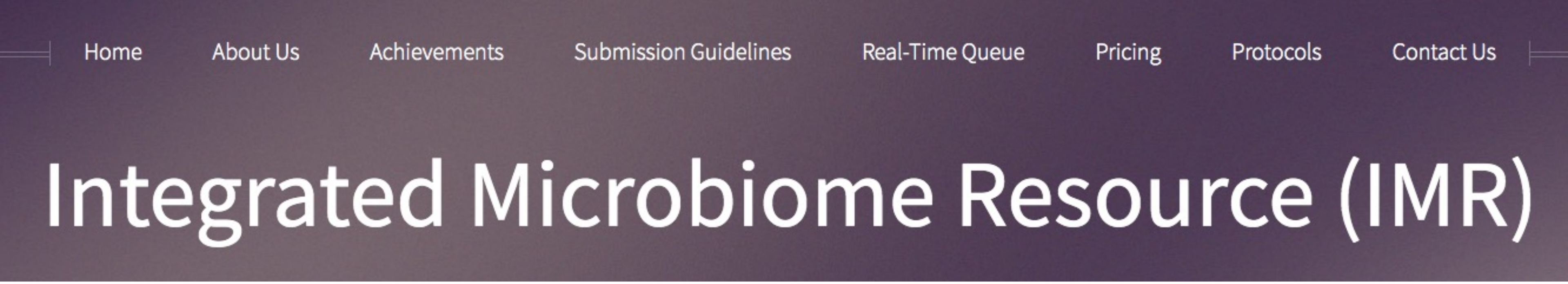
Learning Objectives

- Contrast metagenomic from amplicon sequencing

- Describe general approaches for determining taxonomic composition from metagenomic data

- Describe major steps in constructing and evaluating metagenomic assembled genomes

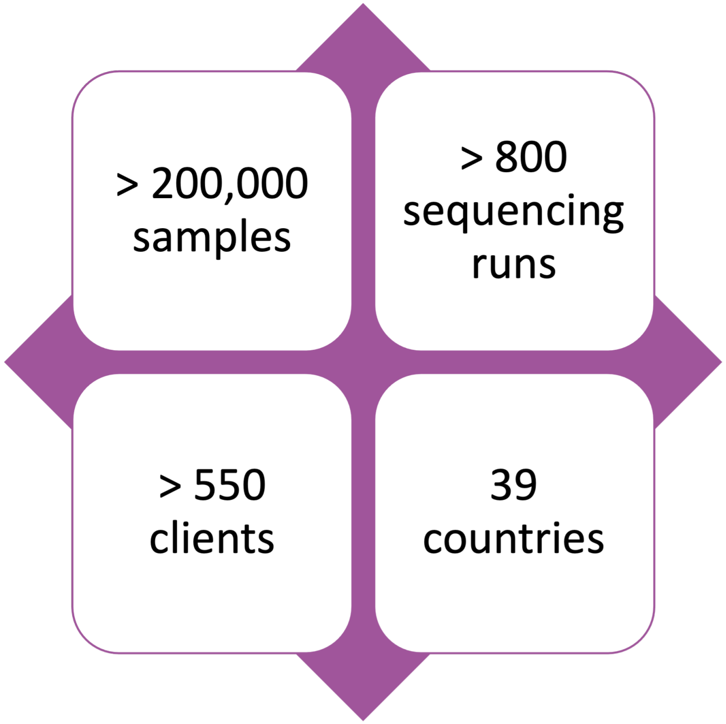

Application &
Resource
Development
Langille
Lab
Collaborative
Health
Projects
Integrated
Microbiome
Resource (IMR)



Integrated Microbiome Resource (IMR)

Sequencing and bioinformatics service for microbiome projects

http://imr.bio

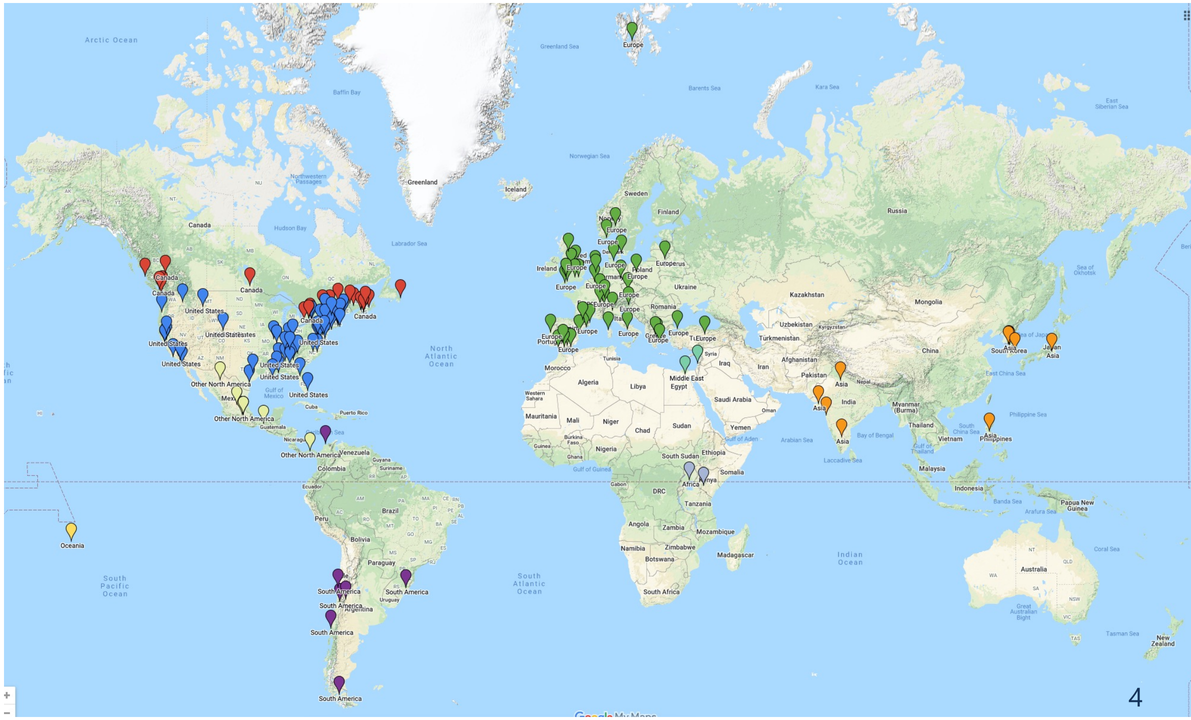


> 200,000 samples

> 800 sequencing runs

> 550 clients

39 countries



Application & Resource Development

Microbiome Helper

https://github.com/LangilleLab/microbiome_helper/

PICRUSt2

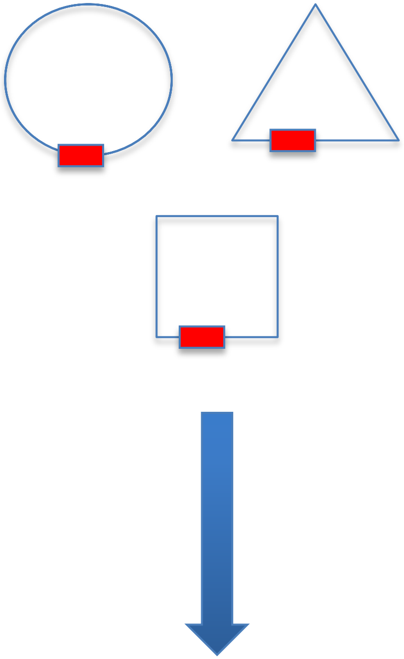
https://github.com/picrust/picrust2/

POMS

https://github.com/gavinmdouglas/POMS

JARVIS

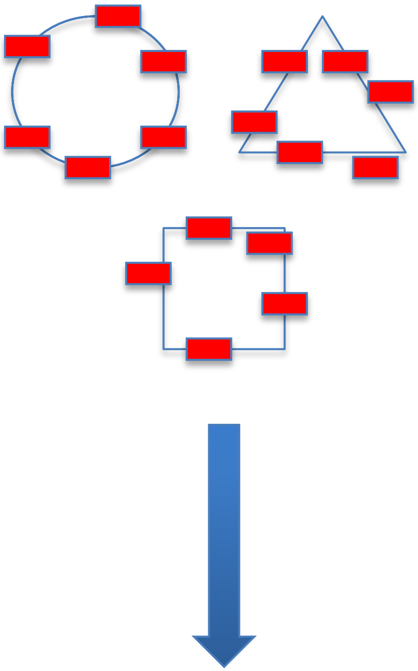
16S rRNA gene sequencing



Who is there?

- 16S: targeted sequencing of the 16S rRNA gene which acts as a marker for identification
  - Well established
  - Relatively inexpensive (~50,000 reads/sample)
  - Only amplifies what you want (no host contamination)

Metagenomics

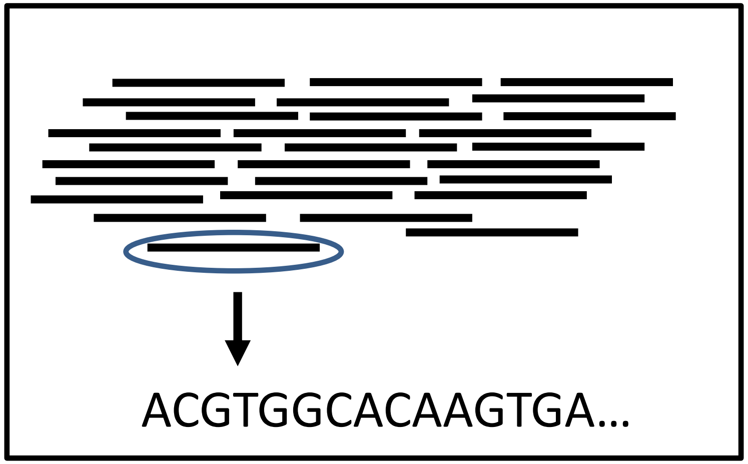


Who is there?
&
What are they doing?

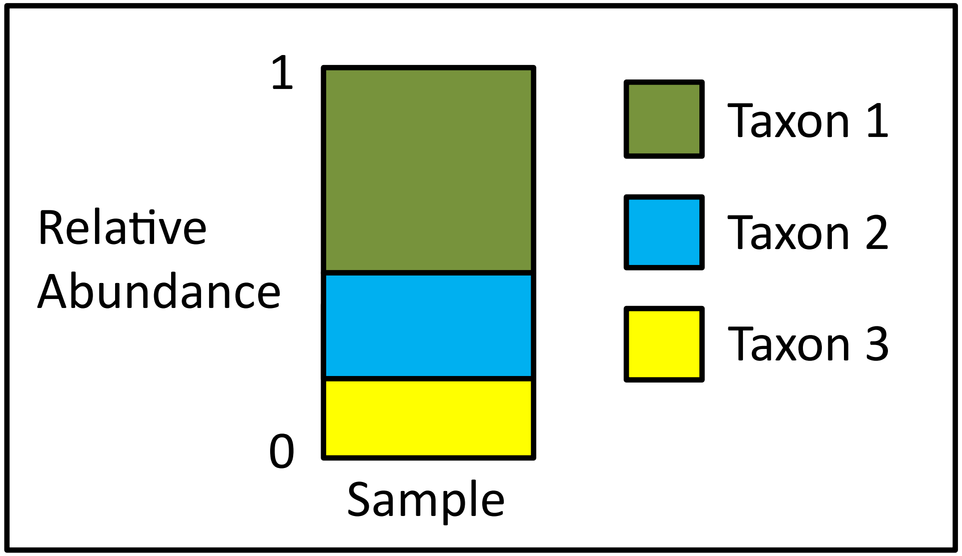
- Metagenomics: sequencing all the DNA in a sample
  - No primer bias
  - Can identify all microbes (bacteria, eukaryotes, viruses)
  - Better taxonomic resolution
  - More expensive (>5-10 million reads/sample)
  - Provides functional information
  - Possibly reconstruct genomes

Taxonomic Profiling

With this raw data:



How do we get this output?



Challenges

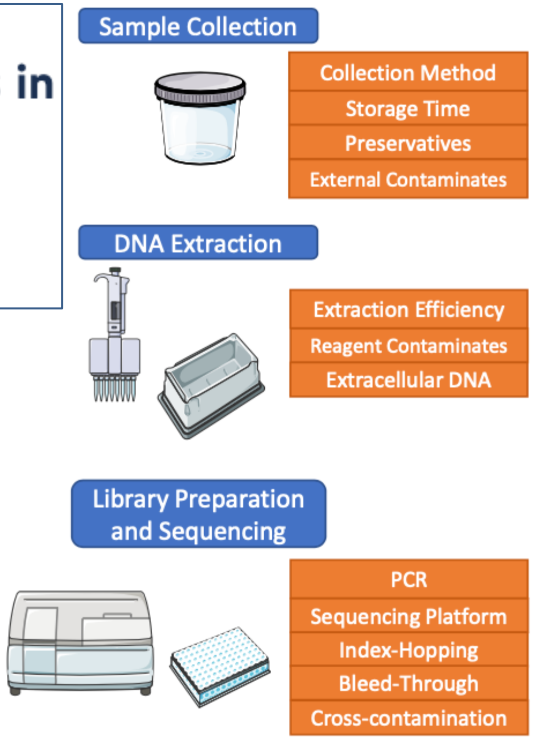
- Reads are randomly assorted
- Reads are usually short (~100-150bp)
- Spotty genome coverage due to sequencing depth
- Lateral gene transfer
- Computational time (Large # reads vs huge databases)
- Let’s not forget about other biases!

Review | Open Access | Published: 18 May 2021

Identifying biases and their potential solutions in human microbiome studies

Jacob T. Nearing, André M. Comeau & Morgan G. I. Langille

Microbiome **9**, Article number: 113 (2021) | Cite this article



Sample Collection

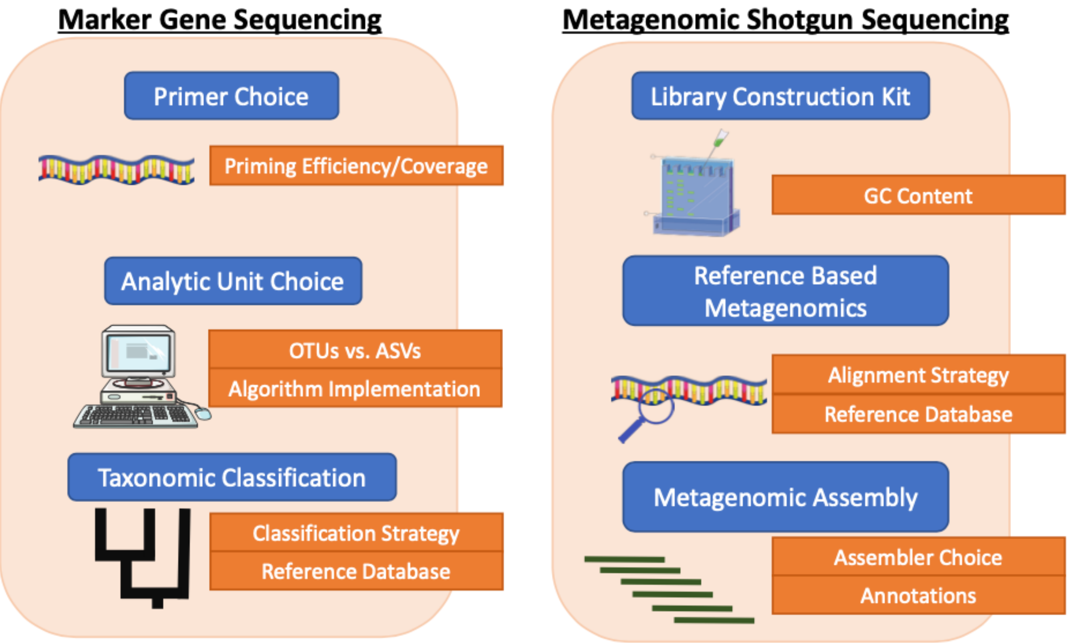
Collection Method
Storage Time
Preservatives
External Contaminates

DNA Extraction

Extraction Efficiency
Reagent Contaminates
Extracellular DNA

Library Preparation and Sequencing

PCR
Sequencing Platform
Index-Hopping
Bleed-Through
Cross-contamination



Marker Gene Sequencing

Primer Choice

Priming Efficiency/Coverage

Analytic Unit Choice

OTUs vs. ASVs
Algorithm Implementation

Taxonomic Classification

Classification Strategy
Reference Database

Metagenomic Shotgun Sequencing

Library Construction Kit

GC Content

Reference Based Metagenomics

Alignment Strategy
Reference Database

Metagenomic Assembly

Assembler Choice
Annotations

Initial bioinformatic processing steps

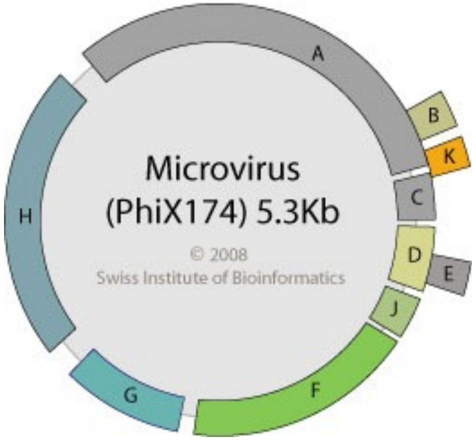
- Many initial steps are similar to 16S studies
- De-multiplexing and lane merging
- Quality filtering
- Stitching paired end reads --> not usually
- **Removal of unwanted host-associated reads**

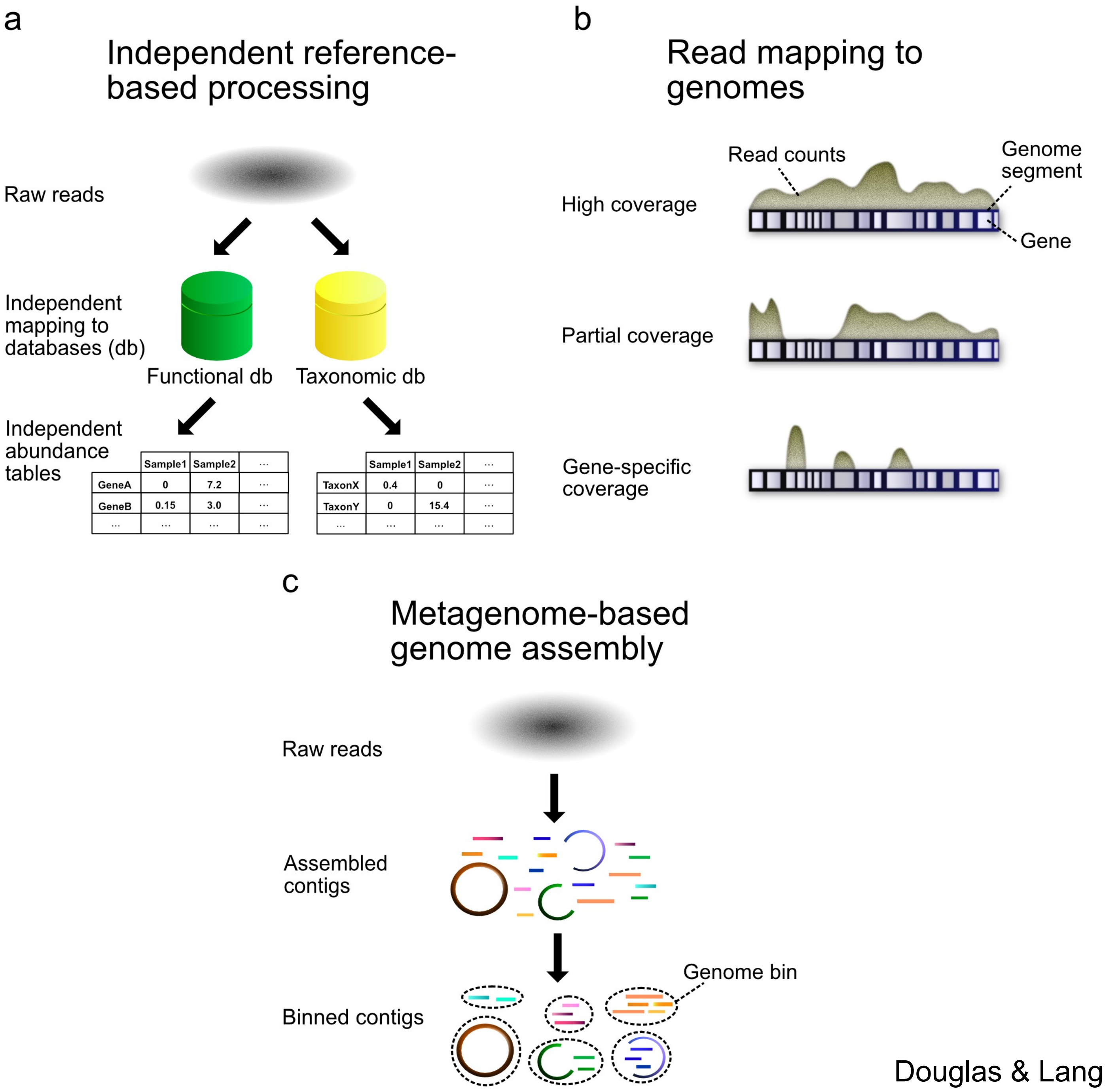
Identifying “contaminant” reads

- Contaminant reads are usually associated with the sampled host (e.g. human, mouse, plant, etc.)

- Typically removed by mapping reads to host reference genome (e.g. bwa, Bowtie2)

- Should filter for Phi X which is used as a sequencing control and is not always removed





a

Independent reference-based processing

b

Read mapping to genomes

c

Metagenome-based genome assembly

	Sample1	Sample2	...
GeneA	0	7.2	...
GeneB	0.15	3.0	...
...	...	...	...

	Sample1	Sample2	...
TaxonX	0.4	0	...
TaxonY	0	15.4	...
...	...	...	...

Douglas & Langille 2021

Reference Based Approaches

- "All reads" approach
  - Attempts to assign taxonomic classification to as many reads as possible
  - Similarity search is computationally demanding
  - May be hard to assign accurate taxonomy to a short read (e.g., repetitive sequence, LGT, no homologs, etc.)

- Marker approaches
  - Uses one or more genome markers to determine the taxonomic composition
  - Only uses a minor subset of the data and thus hard to link to functions downstream
  - Very dependent on choice of markers

Marker Based

- Single Gene
  - Identify and extract reads hitting a single marker gene (e.g. 16S, cpn60, or other “universal” genes)
  - Use existing bioinformatics pipeline (e.g. QIIME, etc.)

- Multiple Gene
  - Several universal genes
    - mOTUs2 (Milanese et al, 2019)
      - Uses 10 universal single copy genes
  - Clade specific markers
    - MetaPhlAn3 (Beghini et al., 2021)

MetaPhlAn3

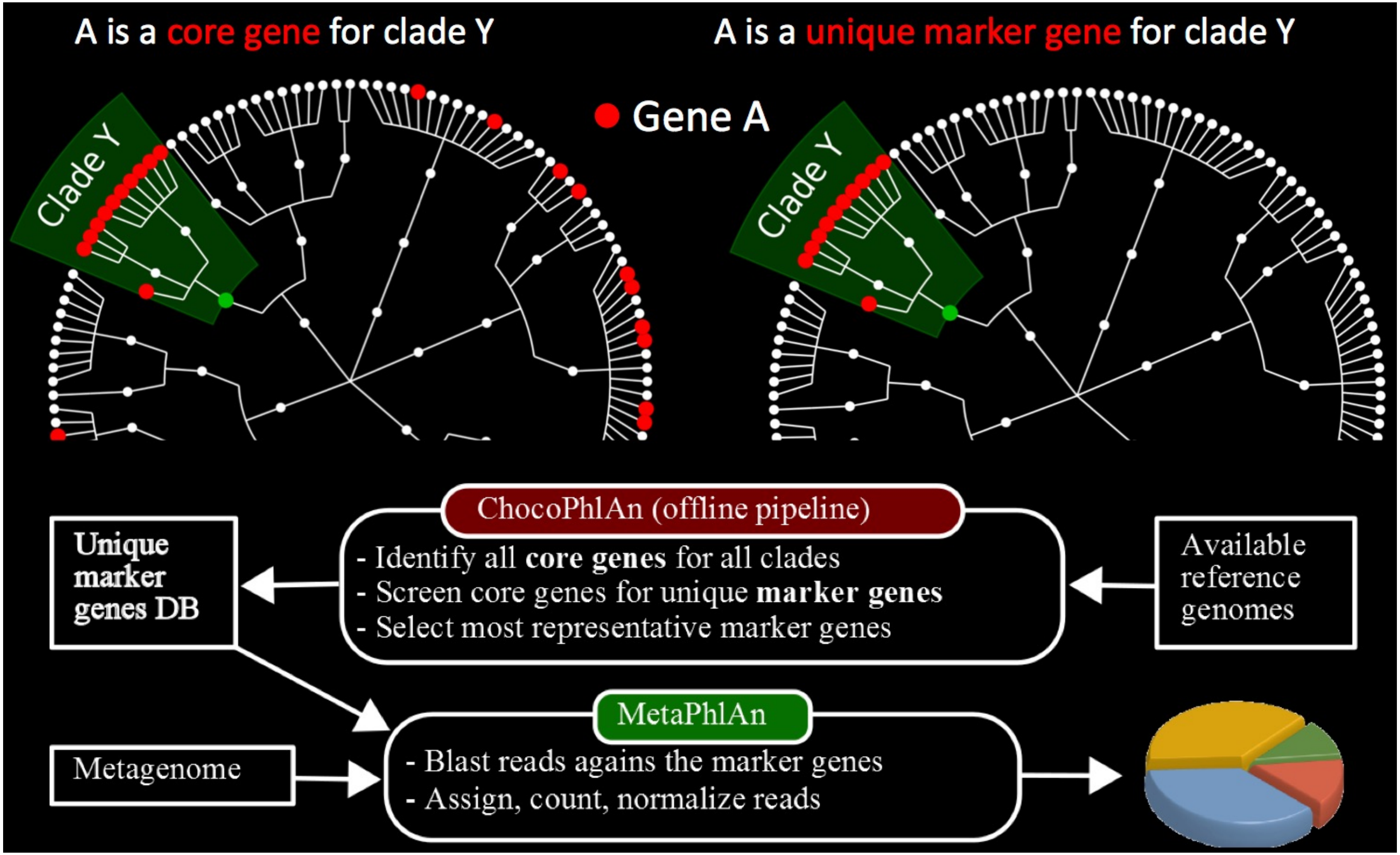
- Uses “clade-specific” gene markers

- Uses ~1.1 million markers derived from ~17,000 genomes

- Can sometimes identify down to the strain level

- Handles millions of reads on a standard computer within a few minutes

MetaPhlAn Marker Selection


A is a core gene for clade Y
A is a unique marker gene for clade Y
● Gene A
Clade Y
Clade Y

Unique marker genes DB

ChocoPhlAn (offline pipeline)

- Identify all **core genes** for all clades
- Screen core genes for unique **marker genes**
- Select most representative marker genes

Available reference genomes

Metagenome

MetaPhlAn

- Blast reads agains the marker genes
- Assign, count, normalize reads

All Reads Approaches

- Kraken/Bracken
- Centrifuge
- Kaiju
- And others!

- Most of these methods use a k-mer based searching solution along with other heuristics to speed up large similarity searches

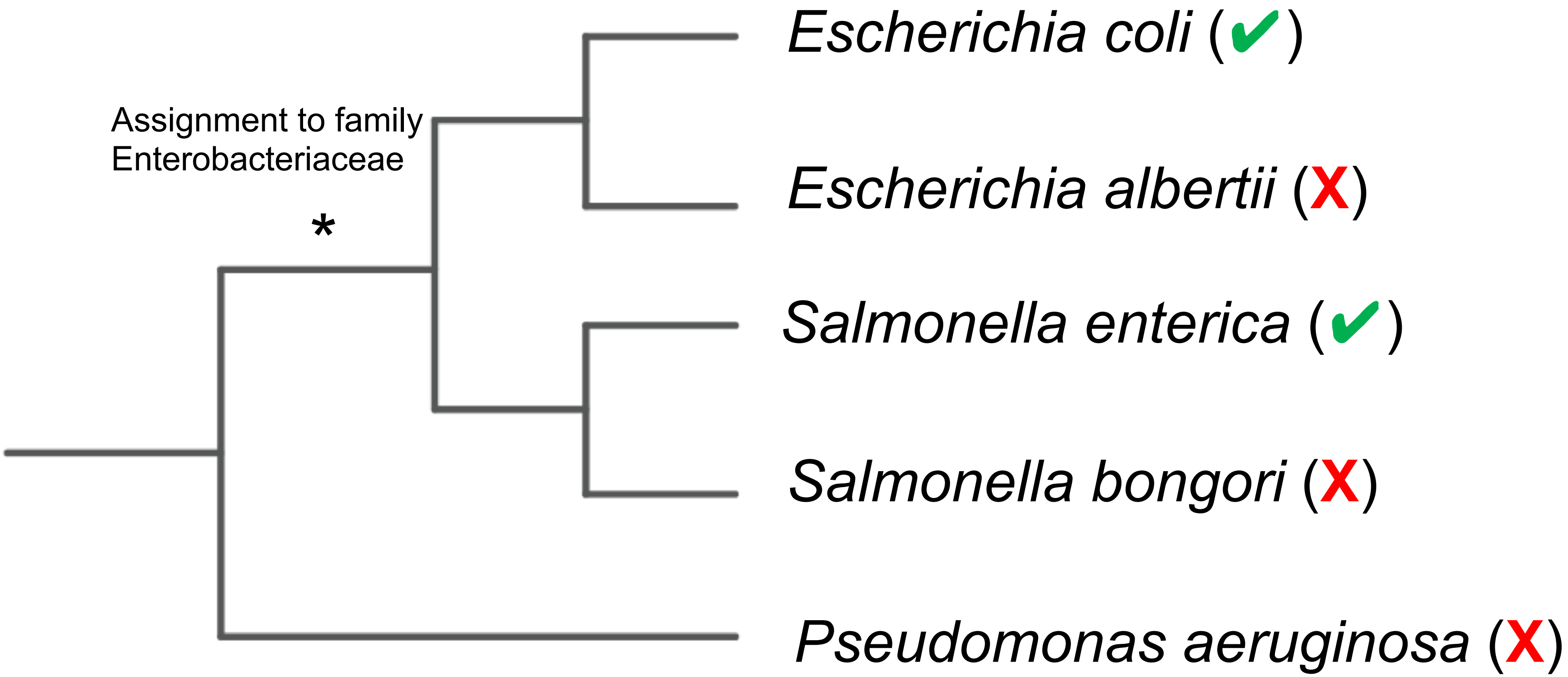
- Many use a lowest common ancestor approach for taxon classification after similarity search

k-mer-based approaches

Sub-sequences of length *k* (*k*-mers)

```
ATCGATCGATCGATCGATCGATCGATC
ATCGA
 TCGAT
  CGATC
   GATCG
    ATCGA
     TCGAT
      CGATC
       GATCG
        ATCGA
         TCGAT
          CGATC
           GATCG
            ATCGA
             TCGAT
```

Lowest Common Ancestor (LCA) Approach



Assignment to family Enterobacteriaceae

*

Escherichia coli (✓)

Escherichia albertii (**X**)

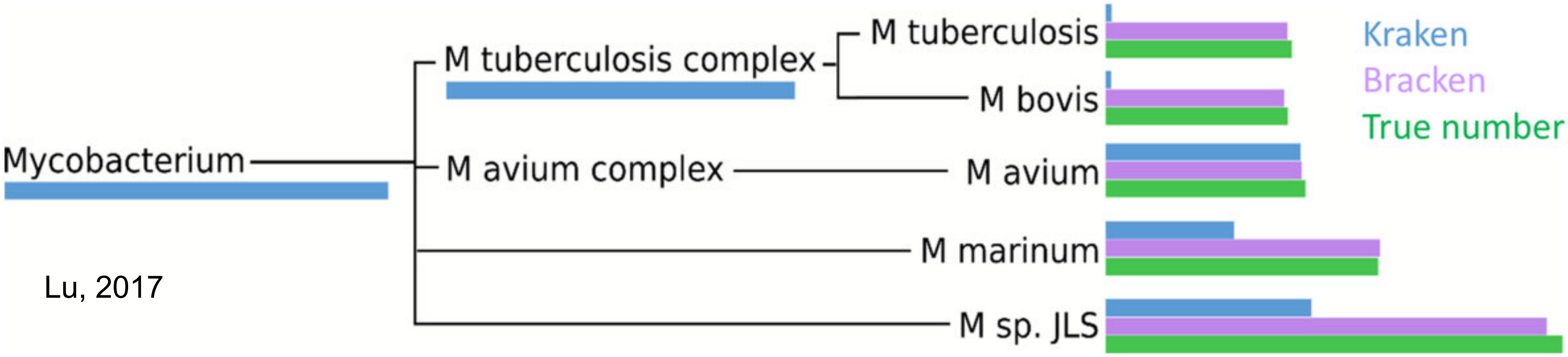
Salmonella enterica (✓)

Salmonella bongori (**X**)

Pseudomonas aeruginosa (**X**)

Kraken & Bracken

- Kraken does the (fast) searching and taxonomy to read
- However, many reads may be placed at a high taxonomic level (e.g. phylum or family) because they are conserved across genomes
- Increasing genomes results in more reads being pushed to higher levels
- Bracken is run after Kraken to improve estimates of species abundance in a sample



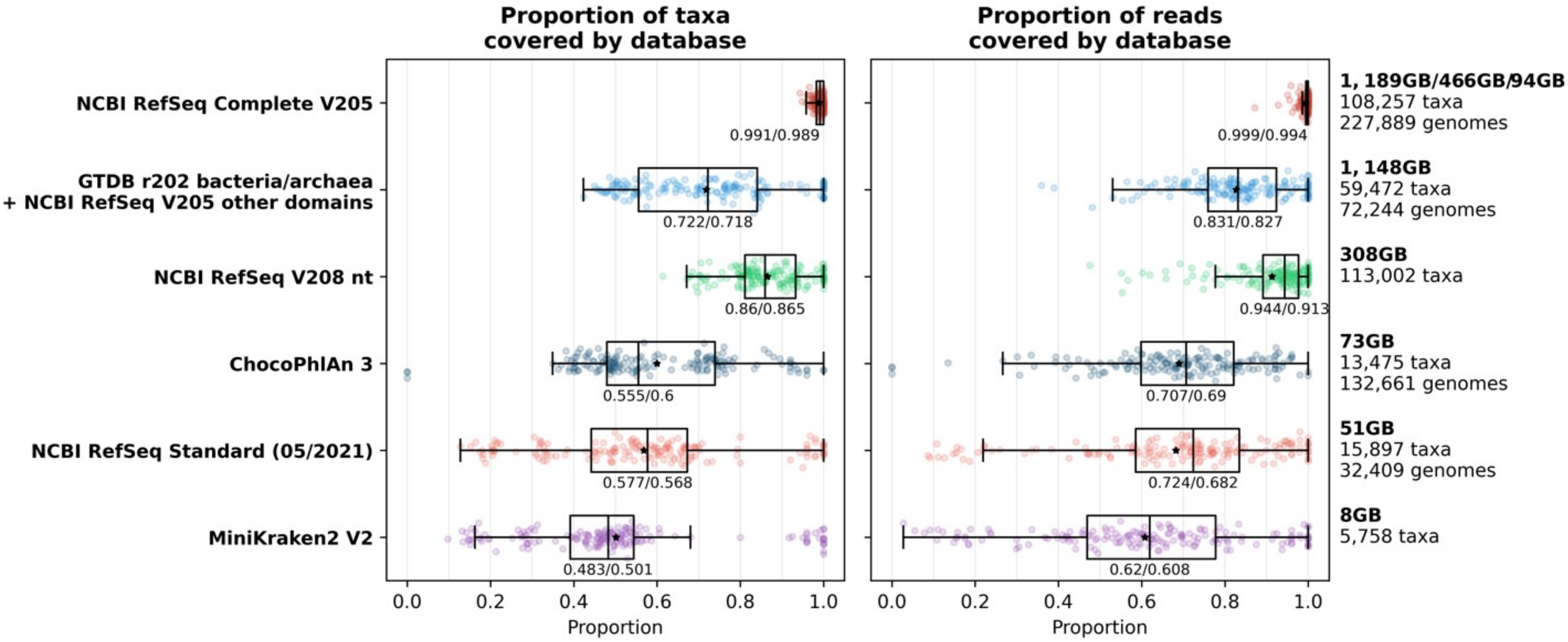
Lu, 2017

Big question: Which is best?

- Difficult to assess comparisons between tools
  - Often different (and often changing) databases
  - Choice of testing dataset (often mock/simulated communities)
  - Choice of tool options/cutoffs
  - Depends who you ask ☺
  - Underlying differences in approaches

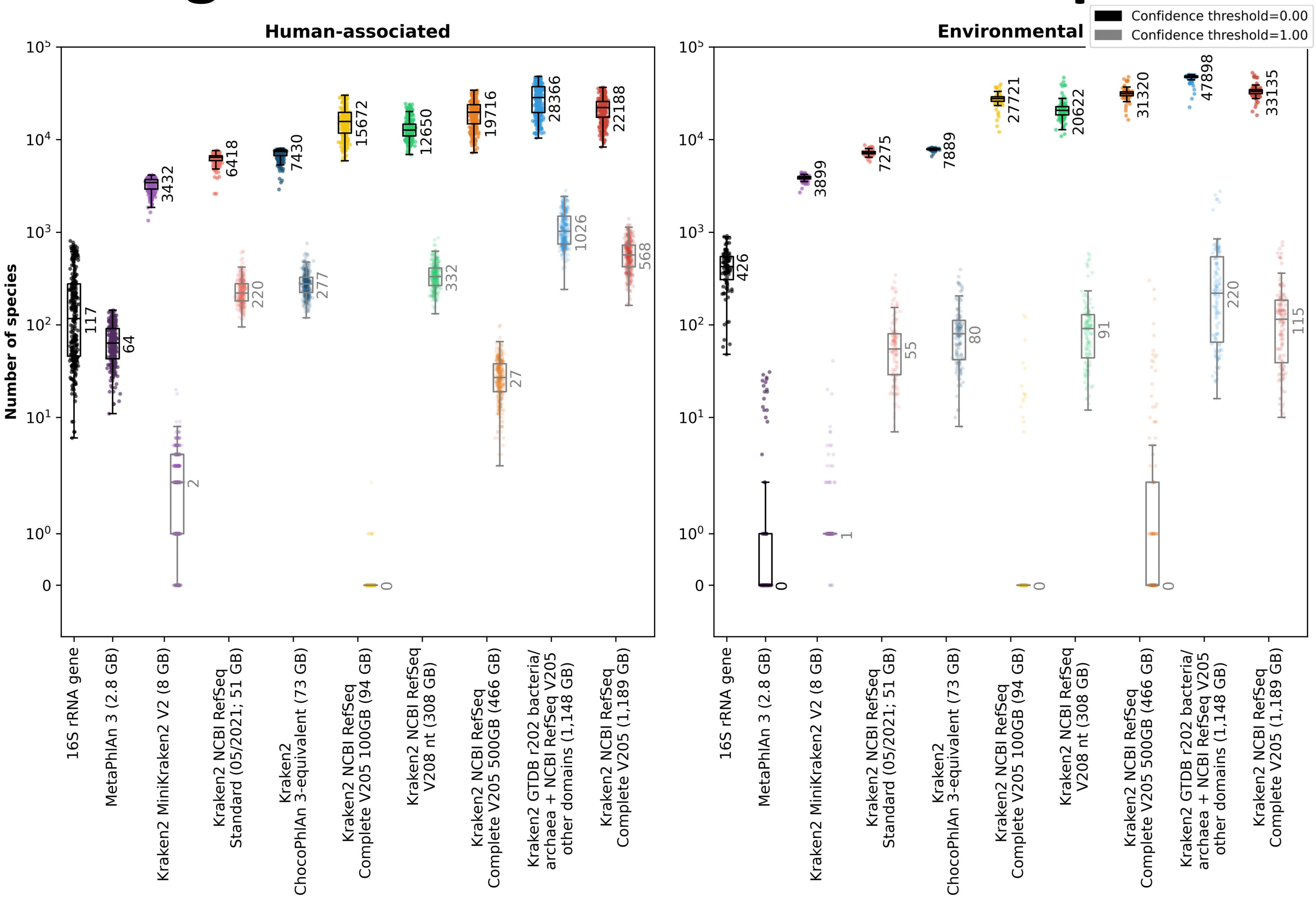
Metaphlan3 vs Kraken 2 Comparison

- Explored the effect of database size and tool parameters

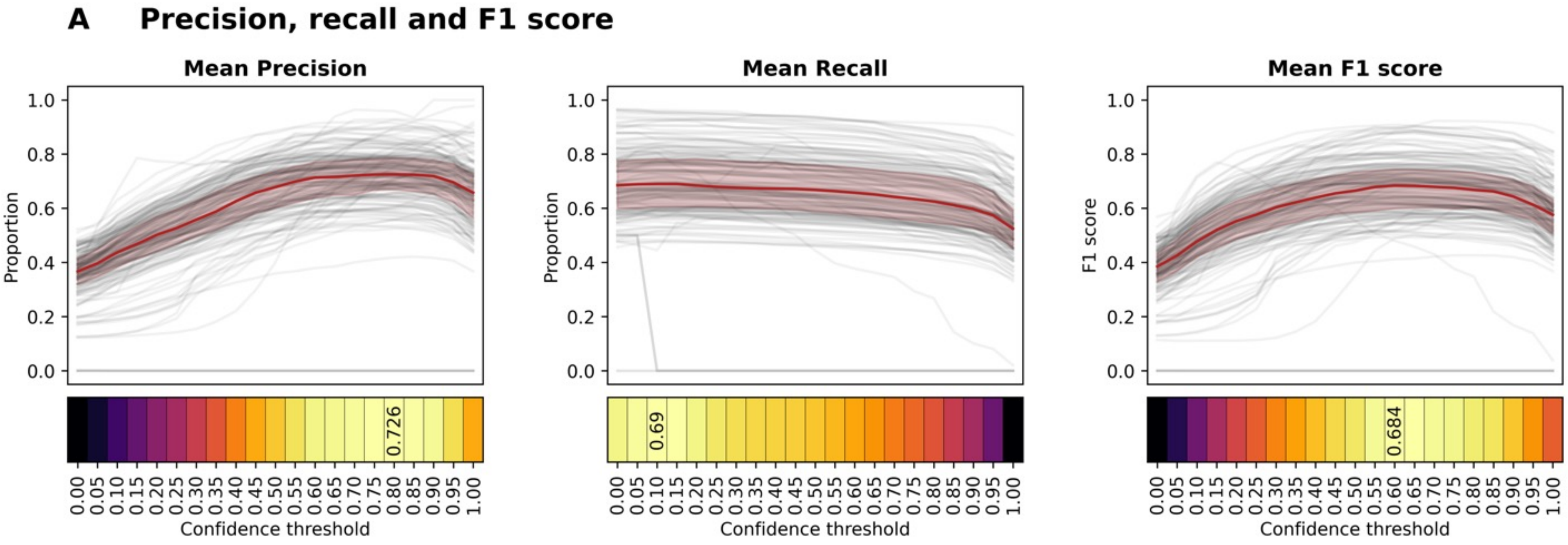


Wright, Comeau & Langille (preprint & in review) From defaults to databases: parameter and database choice dramatically impact the performance of metagenomic taxonomic classification tools

Large differences in number of species

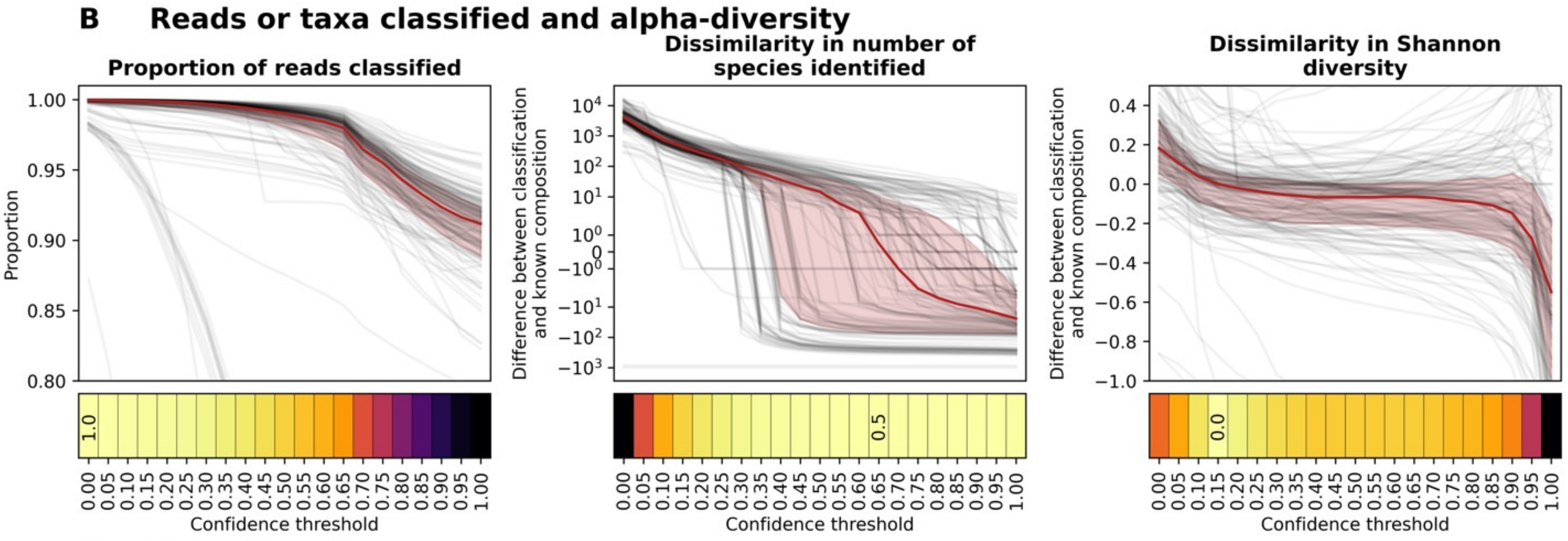


Kraken2 Confidence threshold



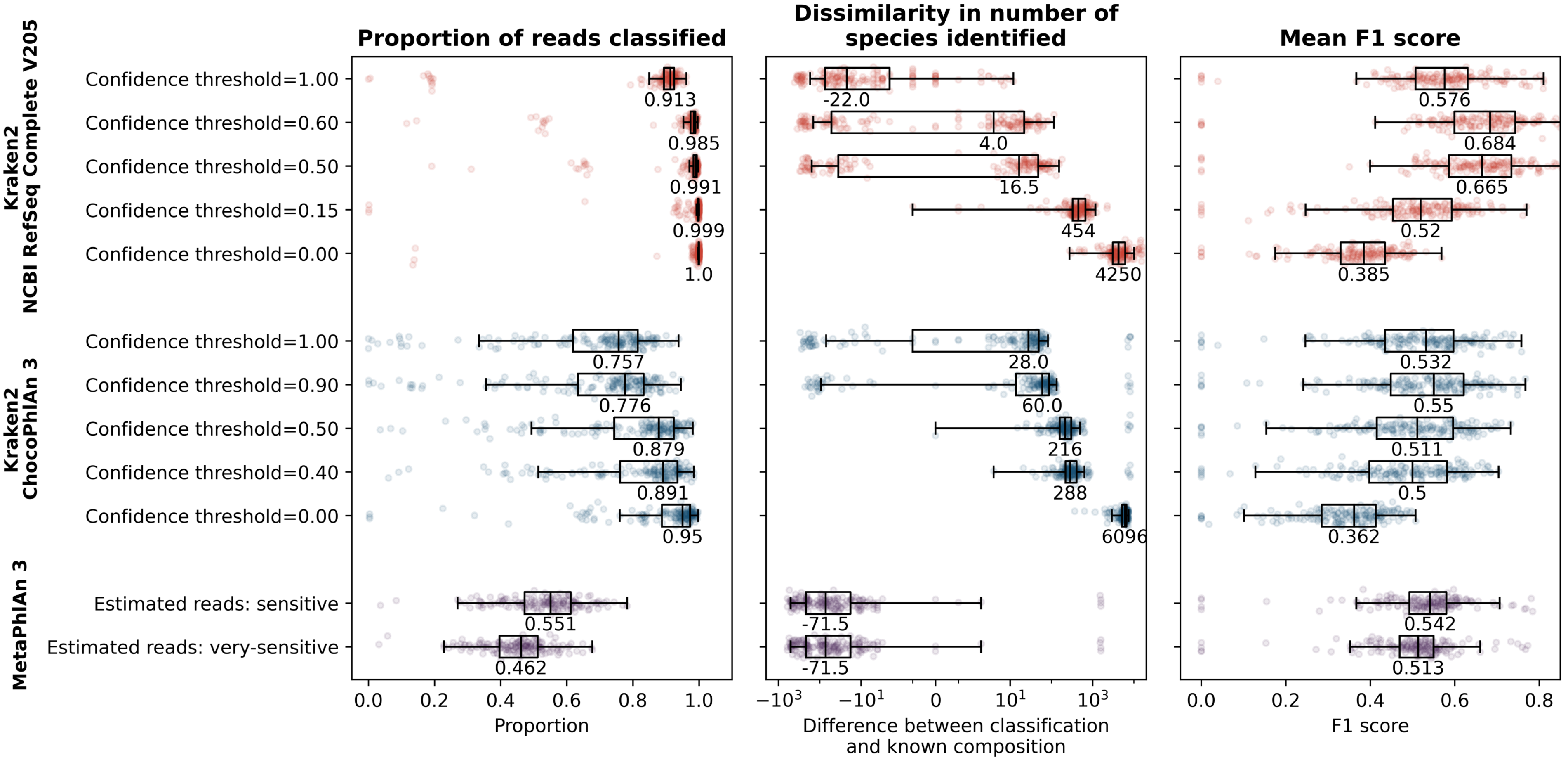
Wright, Comeau & Langille (preprint) From defaults to databases: parameter and database choice dramatically impact the performance of metagenomic taxonomic classification tools

Kraken2 Confidence threshold



Wright, Comeau & Langille (preprint) From defaults to databases: parameter and database choice dramatically impact the performance of metagenomic taxonomic classification tools

Kraken2 vs MetaPhlAn 3



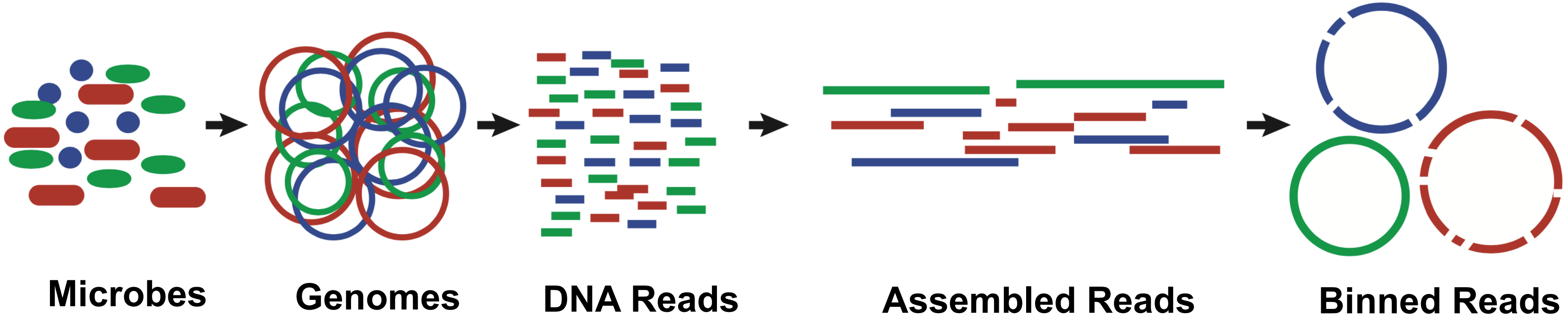
Wright, Comeau & Langille (preprint) From defaults to databases: parameter and database choice dramatically impact the performance of metagenomic taxonomic classification tools

Comparison Summary

- Metaphlan3
  - Fast & low computational requirements,
  - Simple bioinformatic setup (default db and parameters are good)
  - Good for human microbiome studies
  - Good precision (at the cost of some recall)

- Kraken2
  - Good for human AND environmental microbiome studies
  - Confidence cutoff should be changed from default (~0.5)
  - Use as big a database as your computational resources allow (database size equates to amount of memory required)

Metagenomic assembled genomes (MAGs)



Credit to **Dr. Laura A Hug** @ University of Waterloo, for slides, images, and content in this section

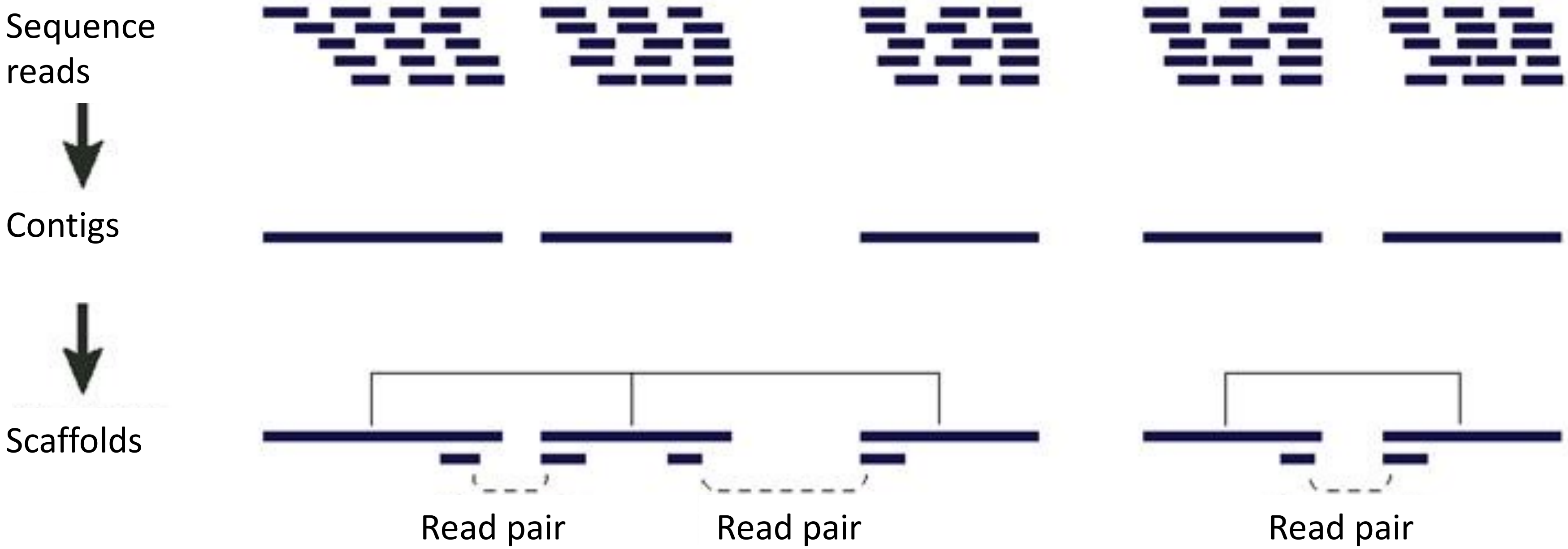
Assembly

- Assembly is the process of generating longer sequence fragments based on read overlaps

- Sequencing strategies and assembly approaches are closely linked
  - Short reads
  - Long reads
  - Linked reads (i.e. 10X)

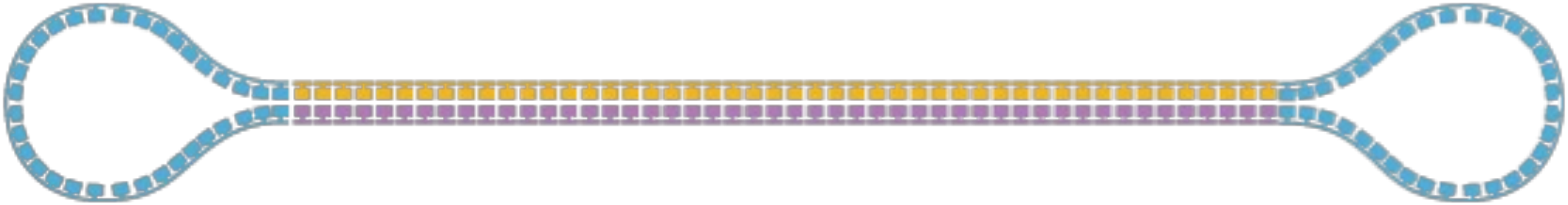
- Many assembly methods (MetaSpades, MEGAHIT, etc,)

Assembling contigs and scaffolds using paired-end reads



Long Reads

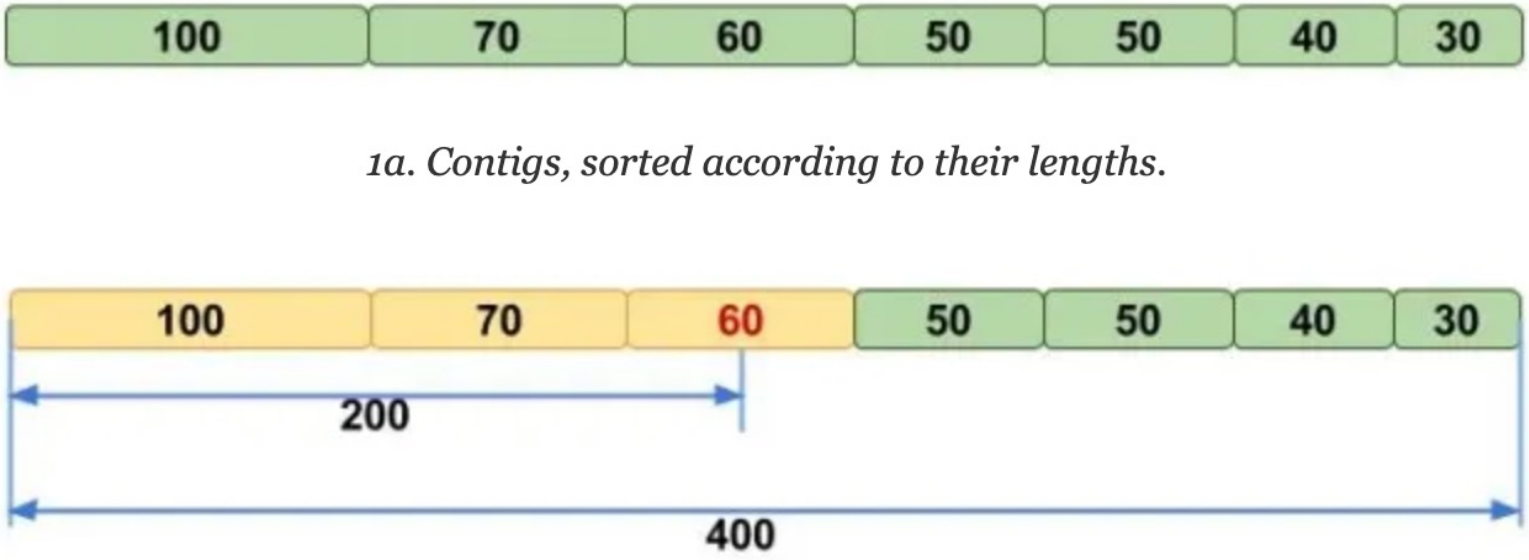
- Long read sequencing becoming increasingly popular
- Two approaches
  - Oxford Nanopore (MinIon)
    - Very long reads (100kb to even mb!)
    - Low infrastructure cost
  - Pacific Biosystems (Pacbio)
    - High throughput
    - Improved accuracy due to “HiFi” reads (e.g. circular consensus sequencing)



Assembly Metrics

- How "good" is my assembly
- MetaQUAST measures assembly quality with several metrics
  - Total length (more is usually better…to a point)
  - Total number of contigs (fewer usually better)
  - Largest contig
  - N50: 50% of the data is within a fragment of this length or greater (bigger is better)

N50



1a. Contigs, sorted according to their lengths.

1b. Calculation of N50 using sorted contigs.

Fig. 1. Example of calculating N50 for a set of seven contigs.
Here N50 equals 60 kbp.

https://www.molecularecologist.com/2017/03/29/whats-n50/

Co-assemble or not?

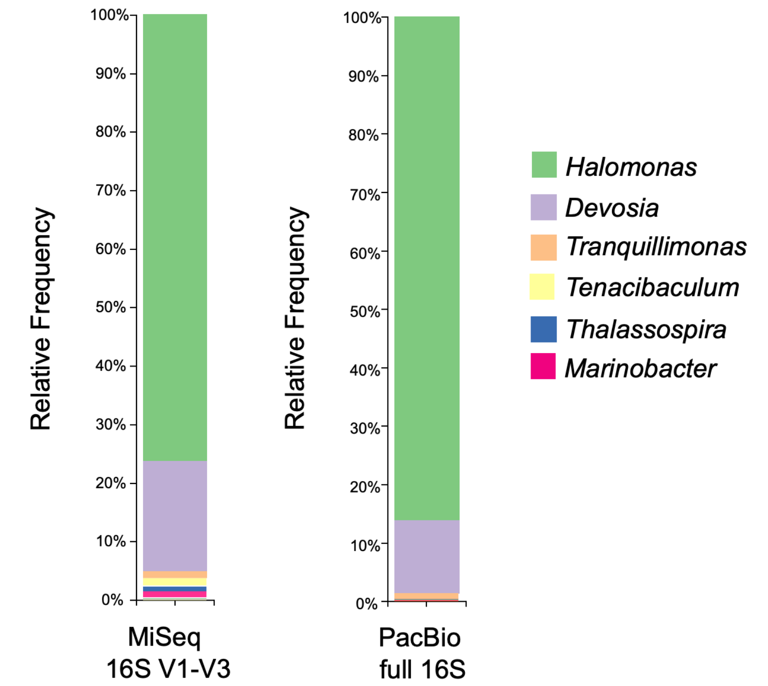
- Co-assembly is the process of combining sequences from multiple samples before assembling

- Advantage
  - More sequence data so likely better assemblies

- Disadvantage
  - Could result in chimeric assemblies

Assembly Example

- Assembly of “simple” bacterial community associated with a unicellular eukaryote



		Short-read assembly		Hybrid assembly	Long-read assembly	
		metaSPAdes	MEGAHIT	metaSPAdes	metaFlye	HiCanu
Whole Assembly	Total number of contigs	2,357	2,301	605	46	107
	Total Length (Mbp)	21.6	21.5	23.8	23.0	24.5
	Contig N50 (bp)	49,951	39,388	277,084	4,078,445	4,095,409
	Largest contig (bp)	669,622	460,209	2,397,197	4,565,899	4,572,073

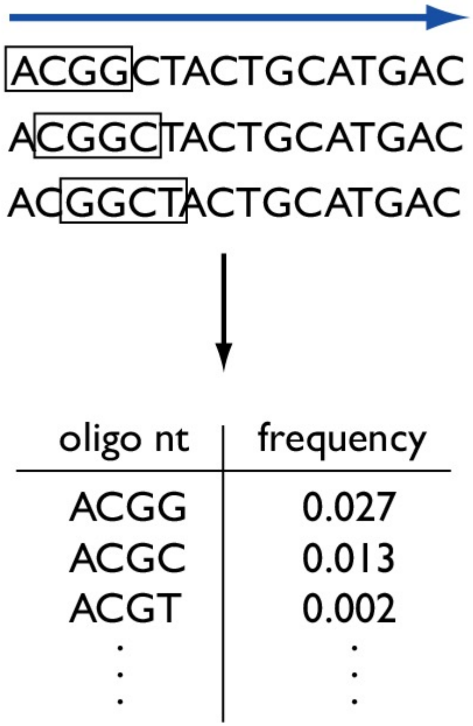
(Filloramo, unpublished)

Binning

- Binning
  - Group (or bin) assembled fragments back into their original genome
  - Generate population-level draft genomes
  - Called metagenome assembled genomes (MAGs)

- Binning methods use one or more of the following characteristics:
  - Nucleotide Composition
  - Phylogenetic affiliation of genes
  - Coverage information

Binning



ACGGCTACTGCATGAC

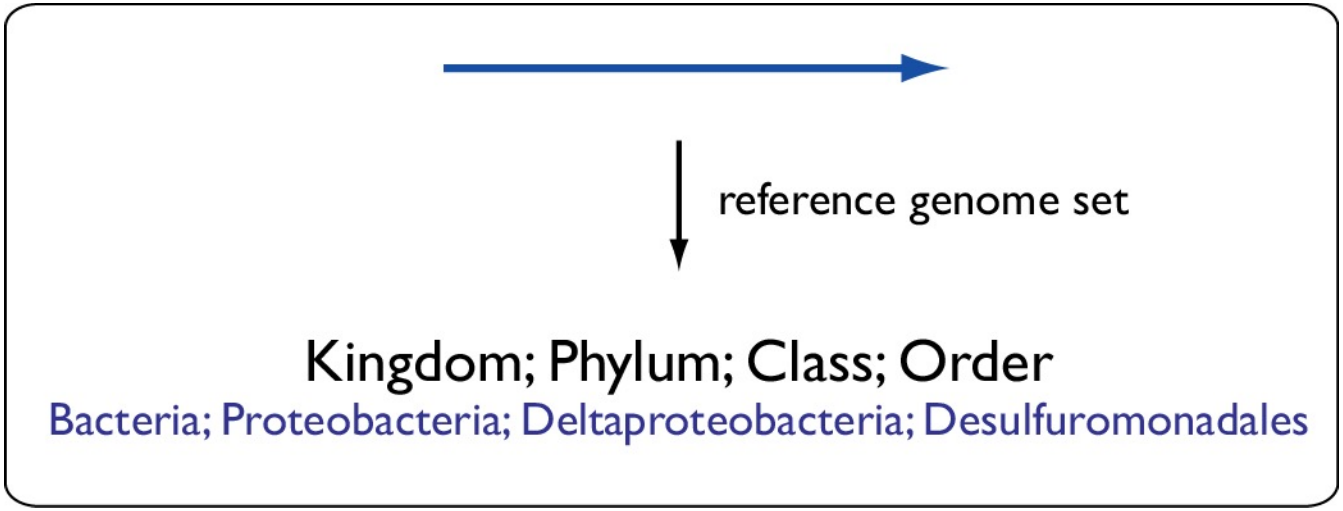
ACGGCTACTGCATGAC

ACGGCTACTGCATGAC

oligo nt	frequency
ACGG	0.027
ACGC	0.013
ACGT	0.002
⋮	⋮

Nucleotide composition

Binning



reference genome set

Kingdom; Phylum; Class; Order

Bacteria; Proteobacteria; Deltaproteobacteria; Desulfuromonadales

Phylogenetic affiliation of scaffolds and/or genes

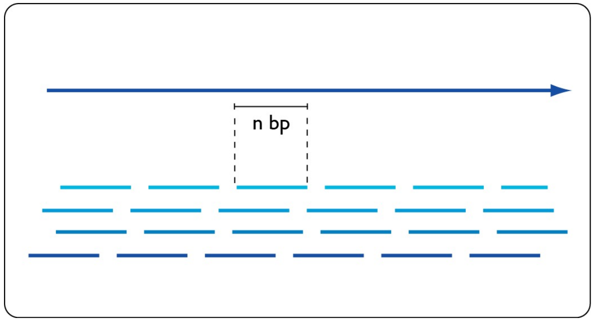
scaffold 1

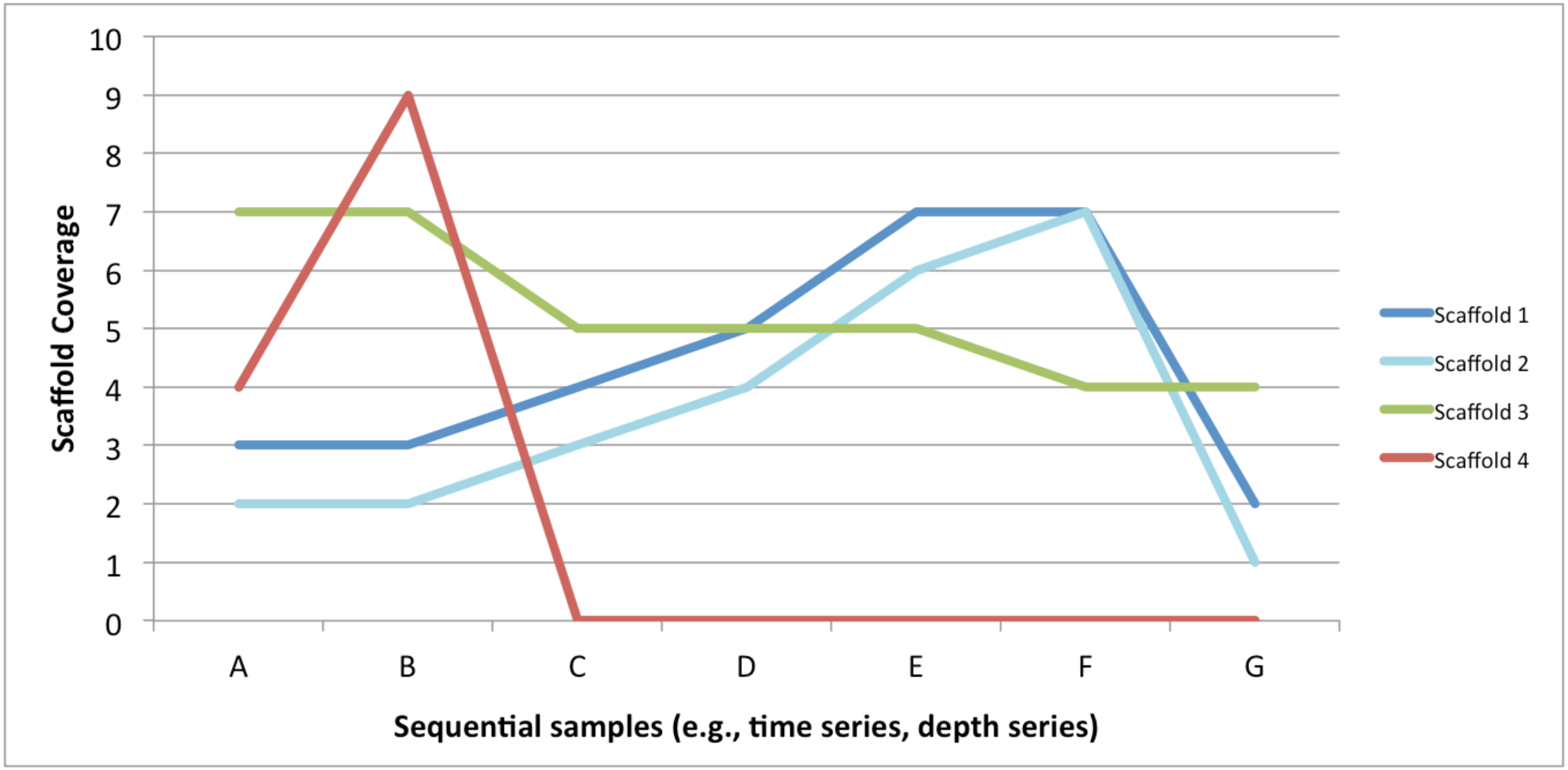
gene A	*Geobacter*
gene B	*Geobacter*
gene C	Deltaproteobacteria
gene D	*Geobacter*

scaffold 2

gene A	Firmicutes
gene B	Chloroflexi
gene C	no hit
gene D	Cyanobacteria

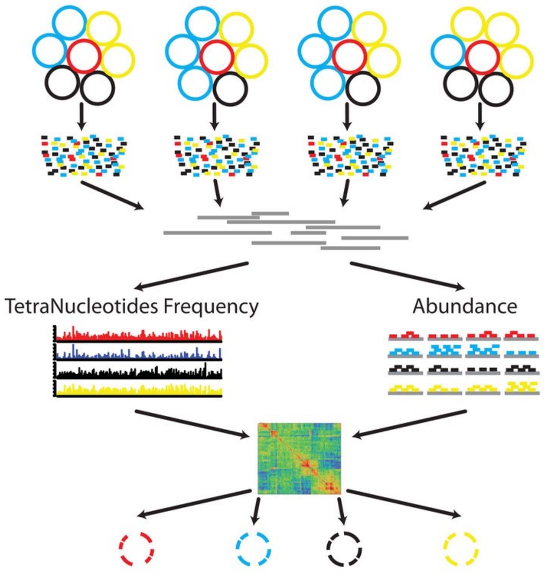
Binning serial samples



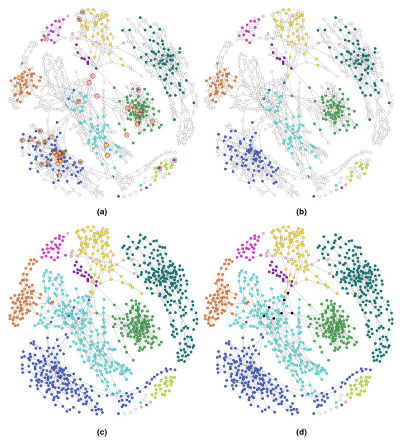


Binning Tools

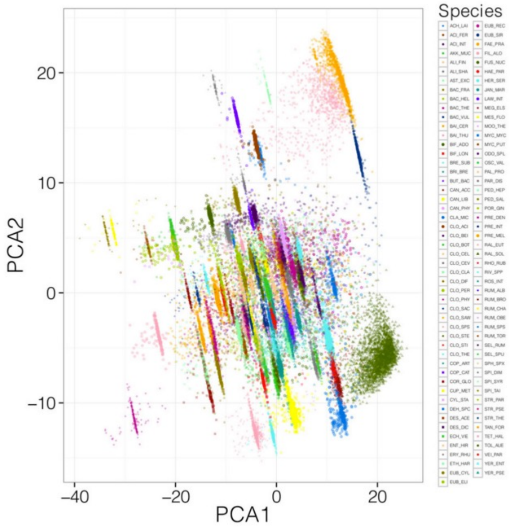
MetaBat



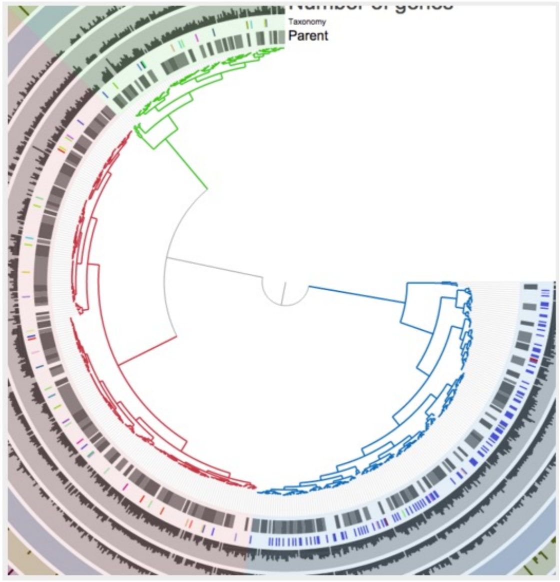
MaxBin2



CONCOCT



Anvi'o



and more…

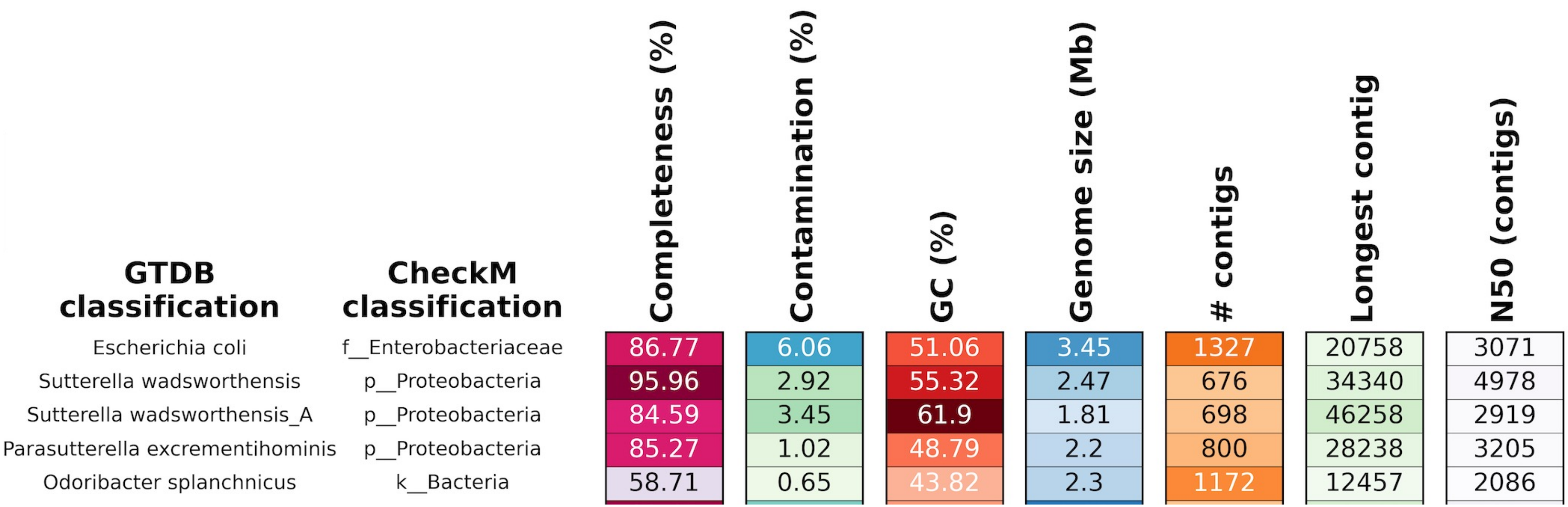
MAG Quality

- Assessing MAG quality is essential!
- Most popular approach is to use single-copy genes
- Completeness
  - Identifies the percentage of single copy genes present in your bin
- Redundancy/Contamination
  - An approximation of what portion of genome is in more than one copy which suggests redundancy

What is this MAG?

- Several approaches to assign taxonomy to each bin
- Approach depends on novelty of the organism and time you want to spend
- Good balance of throughput and approach GTDBtk
- Genome Taxonomy Database (tool kit)

MAG Quality Examples



GTDB classification	CheckM classification	Completeness (%)	Contamination (%)	GC (%)	Genome size (Mb)	# contigs	Longest contig	N50 (contigs)
Escherichia coli	f__Enterobacteriaceae	86.77	6.06	51.06	3.45	1327	20758	3071
Sutterella wadsworthensis	p__Proteobacteria	95.96	2.92	55.32	2.47	676	34340	4978
Sutterella wadsworthensis_A	p__Proteobacteria	84.59	3.45	61.9	1.81	698	46258	2919
Parasutterella excrementihominis	p__Proteobacteria	85.27	1.02	48.79	2.2	800	28238	3205
Odoribacter splanchnicus	k__Bacteria	58.71	0.65	43.82	2.3	1172	12457	2086

Questions?