Canadian Bioinformatics Workshops

www.bioinformatics.ca

bioinformaticsdotca.github.io

Creative Commons

This page is available in the following languages:
Afrikaans български Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE) Euskara Suomeksi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски srpski (latinica) Sotho svenska 中文 華語 (台灣) isiZulu

creative commons

Attribution-Share Alike 2.5 Canada

You are free:

to Share — to copy, distribute and transmit the work

to Remix — to adapt the work

Free Cultural APPROVED FOR Works

Under the following conditions:

Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).

Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.

Disclaimer

Your fair dealing and other rights are in no way affected by the above.
This is a human-readable summary of the Legal Code (the full licence) available in the following languages:
English French

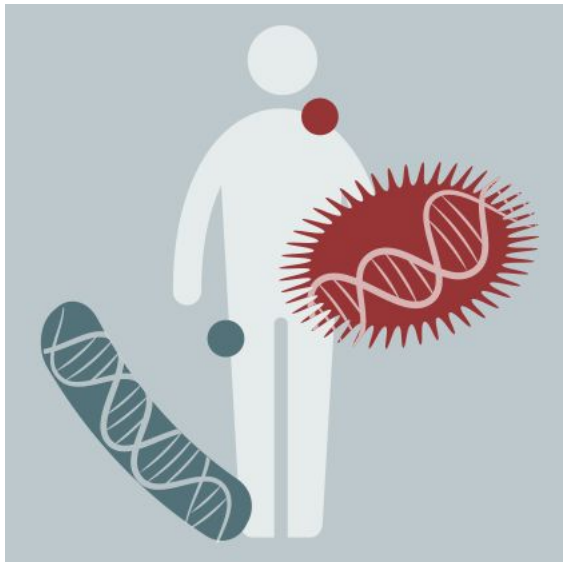
Learn how to distribute your work using this licence

Module 4:Metagenomic Sequencing

Morgan Langille

CBW-IMPACTT Microbiome Analysis

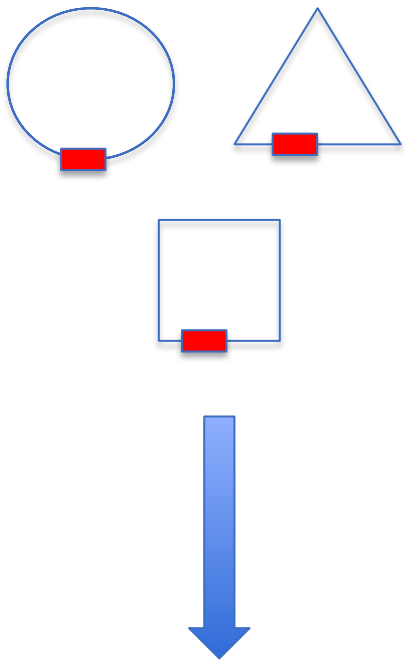
July 5-7, 2023



Learning Objectives

- Contrast metagenomic from amplicon sequencing
- Describe general approaches for determining taxonomic composition from metagenomic data
- Understand microbial functional annotation and tools for annotating metagenomic data with functional labels

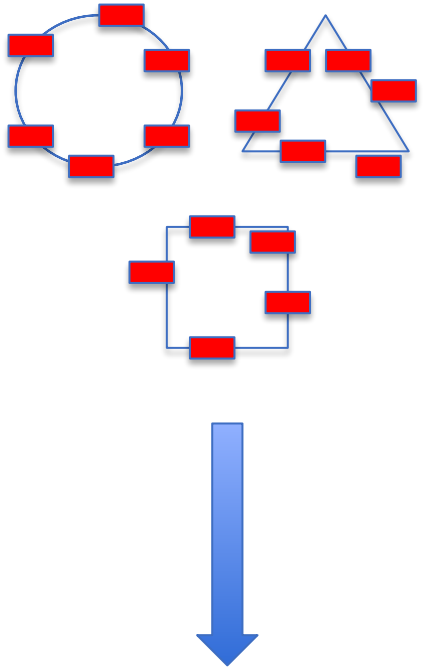
16S rRNA gene sequencing



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

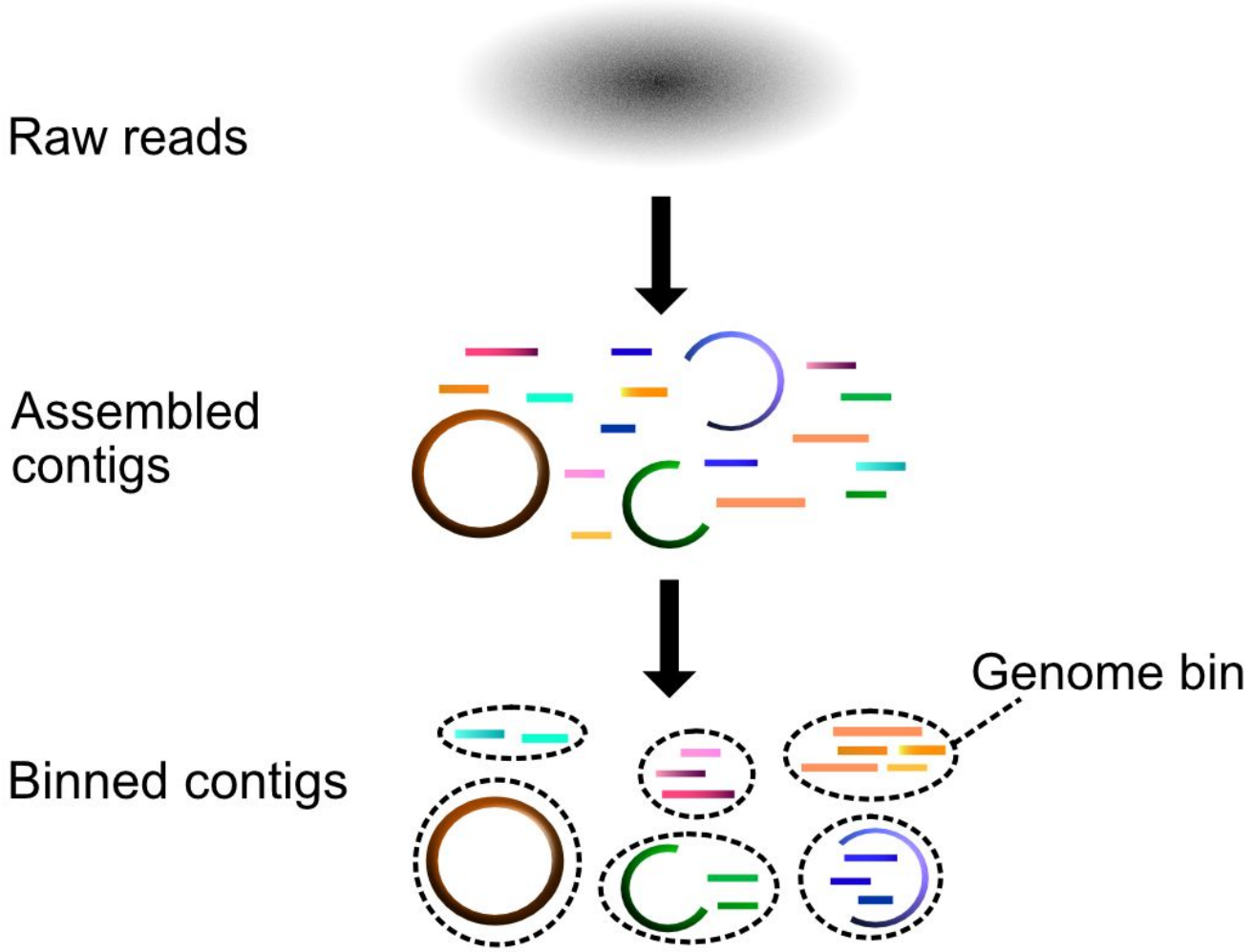


Who is there?
&
What are they doing?

- 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

c

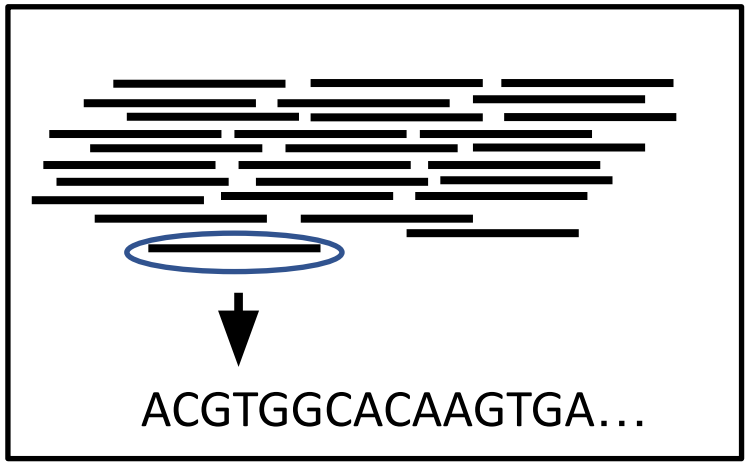
Metagenome-based genome assembly



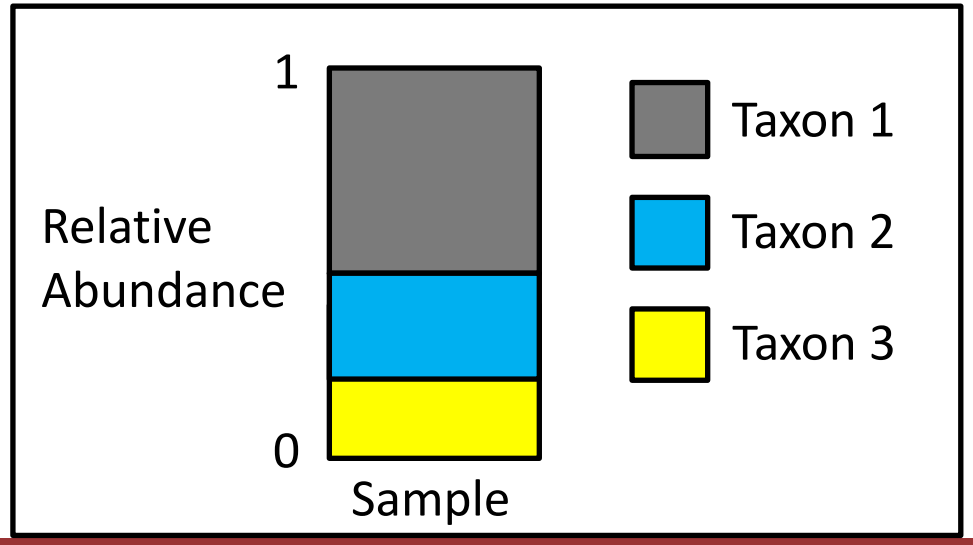
Douglas & Langille 2021

Taxonomic Profiling

With this raw data:



How do we get this output?



Challenges identifying taxa from metagenomics data

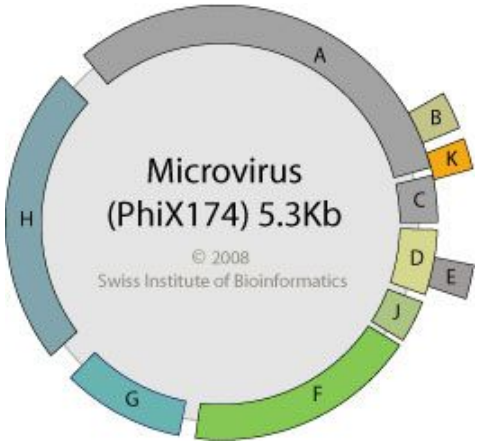
- Reads are randomly assorted
- Reads are *usually* short (~100-150bp)
- Spotty genome coverage due to sequencing depth
- Lateral gene transfer
- Computational time (Large # reads vs huge databases)

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



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
    - MetaPhlAn4 (Blanco-Míguez et al., 2023)

MetaPhlAn4

- Combines ~1 million bacterial and archaeal genomes
  - ~25% from isolate or single cell sequencing
  - ~75% from metagenomically assembled genomes (MAGs)

- Groups genomes into species level genome bins (SGBs, at 5% genomic identity)
  - ~22,000 are known SGBs. ~5,000 unknown SGBs (>5 MAGs)

- 5.1 million unique and core marker genes
  - 10 to 200 markers per genome

- MGS reads are profiled against all markers using Bowtie, filtered, and normalized to produce taxonomic profiles.

- Limited to identifying bacterial and archaea previously identified

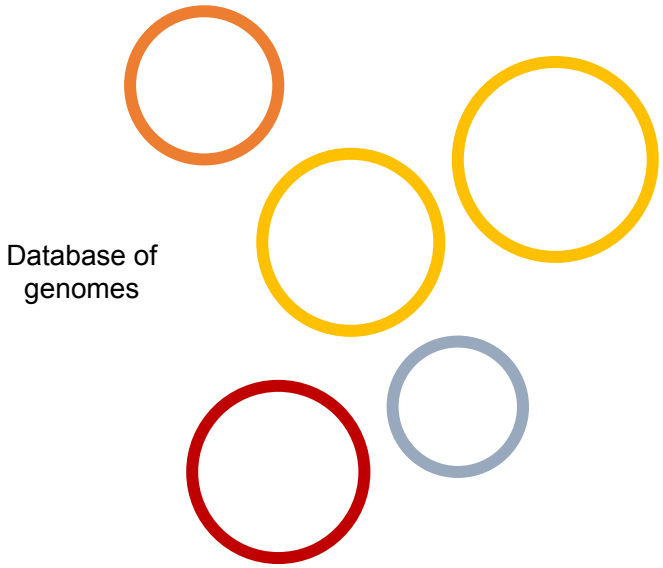
All Reads Approaches

- 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

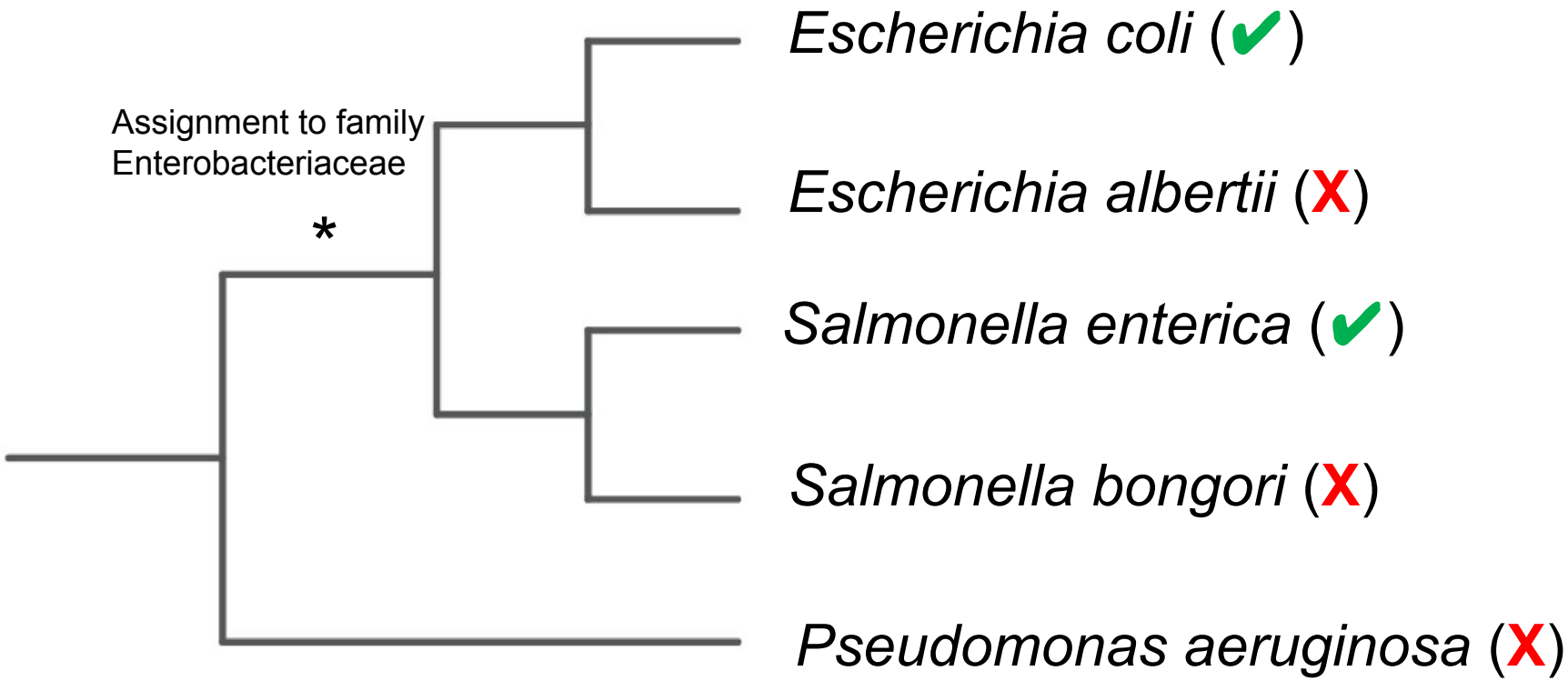
k-mer-based approaches



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

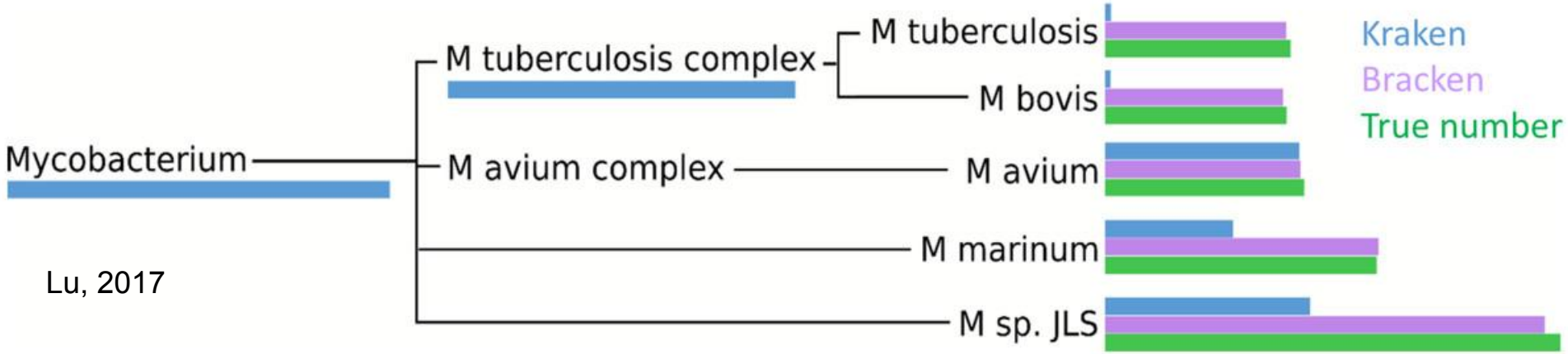
```
ATCGATCGATCGATCGATCGATCGATCGATC
ATCGA
 TCGAT
  CGATC
   GATCG
    ATCGA
     TCGAT
      CGATC
       GATCG
        ATCGA
         TCGAT
          CGATC
           GATCG
            ATCGA
             TCGAT
```

Lowest Common Ancestor (LCA) Approach



Kraken & Bracken

- Kraken does the (fast) searching and assigning taxonomy to reads
- 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

MICROBIAL GENOMICS
Volume 9, Issue 3

Research Article | Open Access

From defaults to databases: parameter and database choice dramatically impact the performance of metagenomic taxonomic classification tools

Robyn J. Wright[1], Andrè M. Comeau[2], Morgan G. I. Langille[1,2]
View Affiliations

Published: 03 March 2023 | https://doi.org/10.1099/mgen.0.000949

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)

Functional Composition

- Taxonomic composition answers “Who is there?”
- Functional composition answers “What are they doing?”
- Metagenomics provides the opportunity to catalog the set of genes from an entire community and compare those differences across samples

What do we mean by function?

- General categories
  - Photosynthesis
  - Nitrogen metabolism
  - Glycolysis

- Specific groups of orthologs
  - Nifh
  - EC: 1.1.1.1 (alchohol dehydrogenase)
  - K00929 (butyrate kinase)

Various Functional Databases

- COG
  - Well known but original classification (not updated since 2003)
- SEED
  - Used by the RAST and MG-RAST systems
- PFAM
  - Focused more on protein domains
- UniRef
  - Has clustering at different levels (e.g. UniRef100, UniRef90, UniRef50)
  - Most comprehensive and is constantly updated
- KEGG
  - Very popular, each entry is well annotated, and often linked into “Modules” or “Pathways”
  - Full access requires a license fee
- MetaCyc
  - Very popular and the primary alternative/replacement for KEGG

Metagenomics Annotation Systems

- Web-based (These all provide functional and taxonomic analysis, plus hosts your data.)
  - MGnify (i.e. EBI Metagenomics Server)
  - MG-RAST
  - IMG/M

- Graphical user interface:
  - MEGAN
    - Provides several visualizations

- Local-based (many more not listed here)
  - Carnelian
    - K-mer based approach, calls genes first with fragGeneScan
  - HUMAnN3
    - Popular and fast
  - Microbiome Helper
    - Custom pipeline (used in lab)

Challenges in Functional Annotation

- Similar challenges to microbial genome annotation
- Partial gene fragments
- Mixed communities
  - Including microbial euks and viruses!
- Large amounts of data! Speed and scalability are very important!
- **Bias from gene lengths**
- **Inferring modules pathways with multiple organisms**

Keys steps of a functional annotation pipeline

- Similarity search approach
  - Mappers like BWA and Bowtie are very fast but limited to DNA space and very similar sequences
  - Protein alignments like DIAMOND and mmSeqs2 are slower but more sensitive in protein space
    - Note that tools like BLAST are usually too slow for any metagenomics annotation
- Database
  - Larger databases are more comprehensive and will annotate larger portions of your samples
  - Smaller databases can be used if focused on well-annotated functions

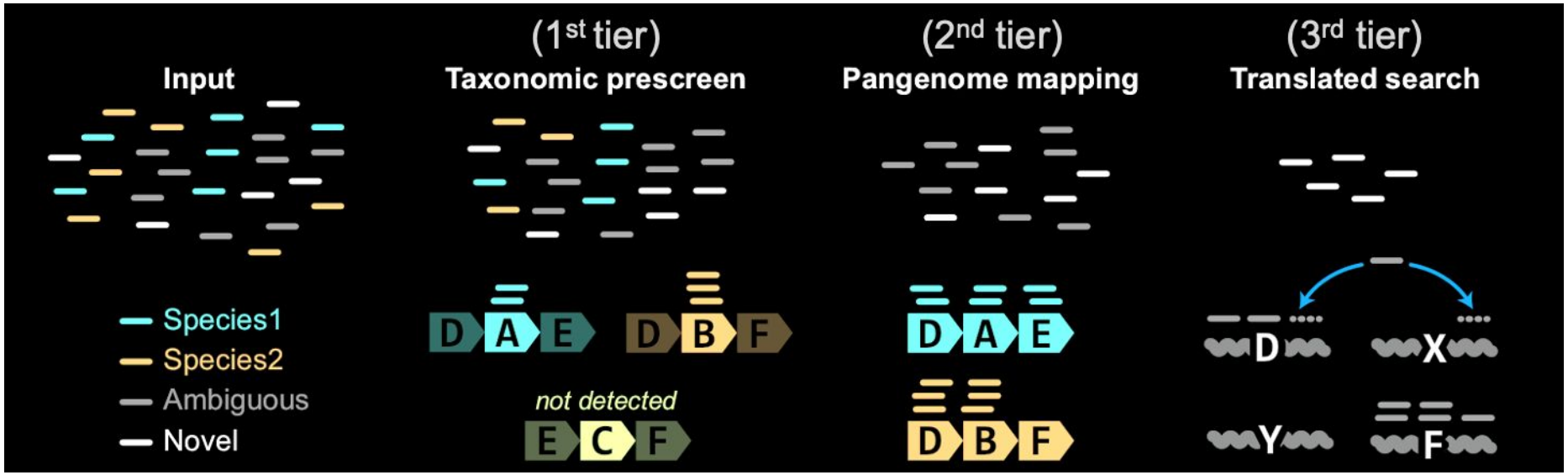
“Normalization”

- Larger genes are more likely to be sequenced than small genes.
- Thus, normalize gene annotations by their length is very common
  - Results are often reported as RPKM (reads per kilobase per million)
- Some methods may also account for similarity % to db, genome size, scaling factor (so not small decimals)

Pathway Inference

- Goal is to reduce spurious pathways
- A KO/EC can map to one or more KEGG/MetaCyc Pathways
  - Just because a gene is found in a pathway doesn't mean that it exists in the community
  - If a pathway has 20 genes and only 2 genes are observed in the community (but at high abundances) what should be the abundance of the pathway?
  - MinPath attempts to estimate the abundance of these pathways and remove spurious noise

Humann



- 1st: Metaphlan3 used to identify species
- 2nd: Reads are mapped with Bowtie to pangenomes of species identified by Metaphlan
- 3rd: Left over reads are translated searched (nucleotide vs protein database) using DIAMOND

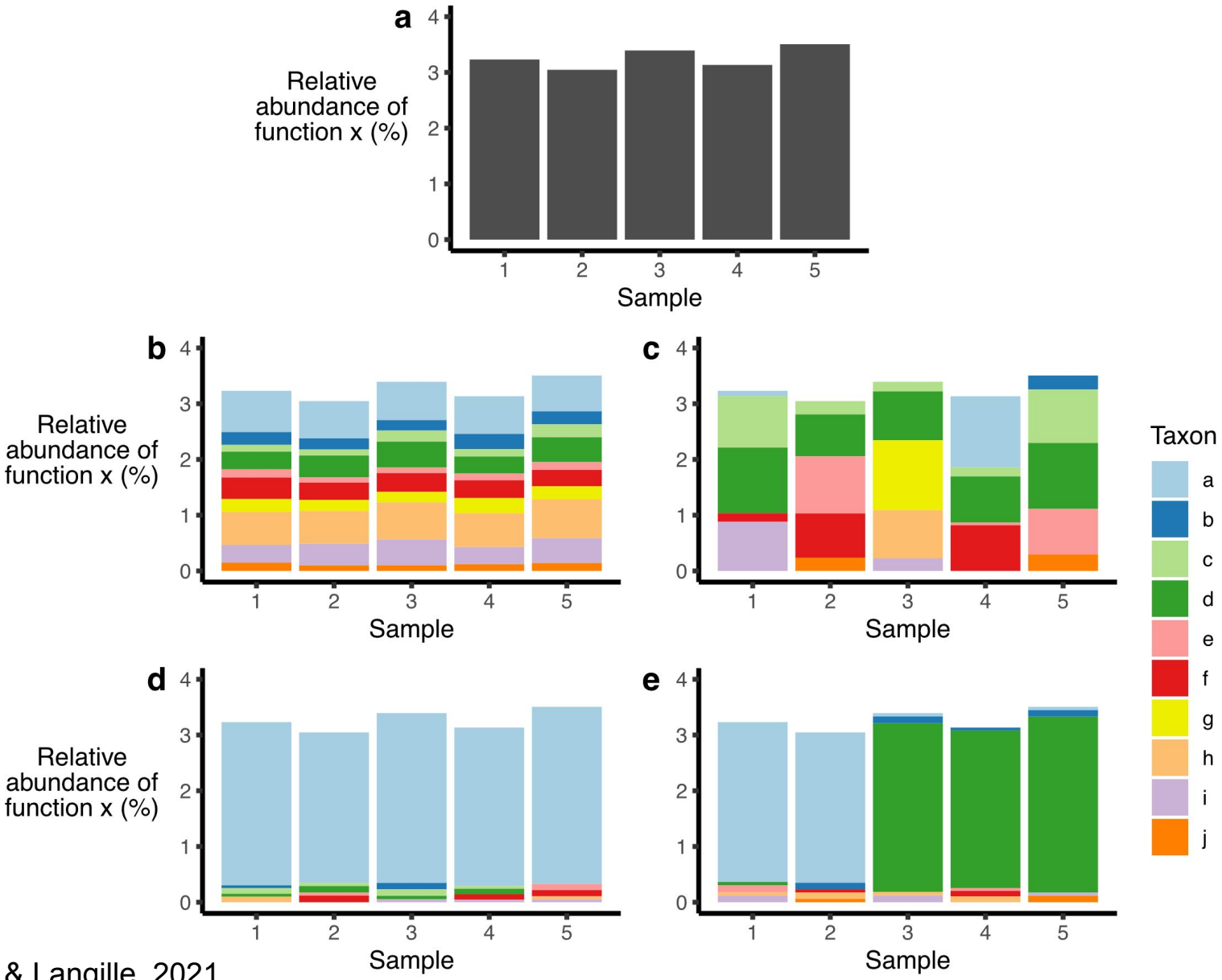
HUMAnN2: stratified output

UniRef gene cluster	Gene name	Total gene abundance (RPK)
UniRef90_R6K3Z5	IMP dehydrogenase	600.95
	Bacteroides_caccae	234.76
	Bacteroides_dorei	107.38
	Bacteroides_ovatus	92.18
	Bacteroides_stercoris	83.95
	Bacteroides_vulgatus	57.27
	unclassified	25.41

Per-species unclassified

MetaCyc pathway	abundance	coverage
PWY-7221: GTP biosynthesis	200.35	1
	120.23	1
	11.12	0

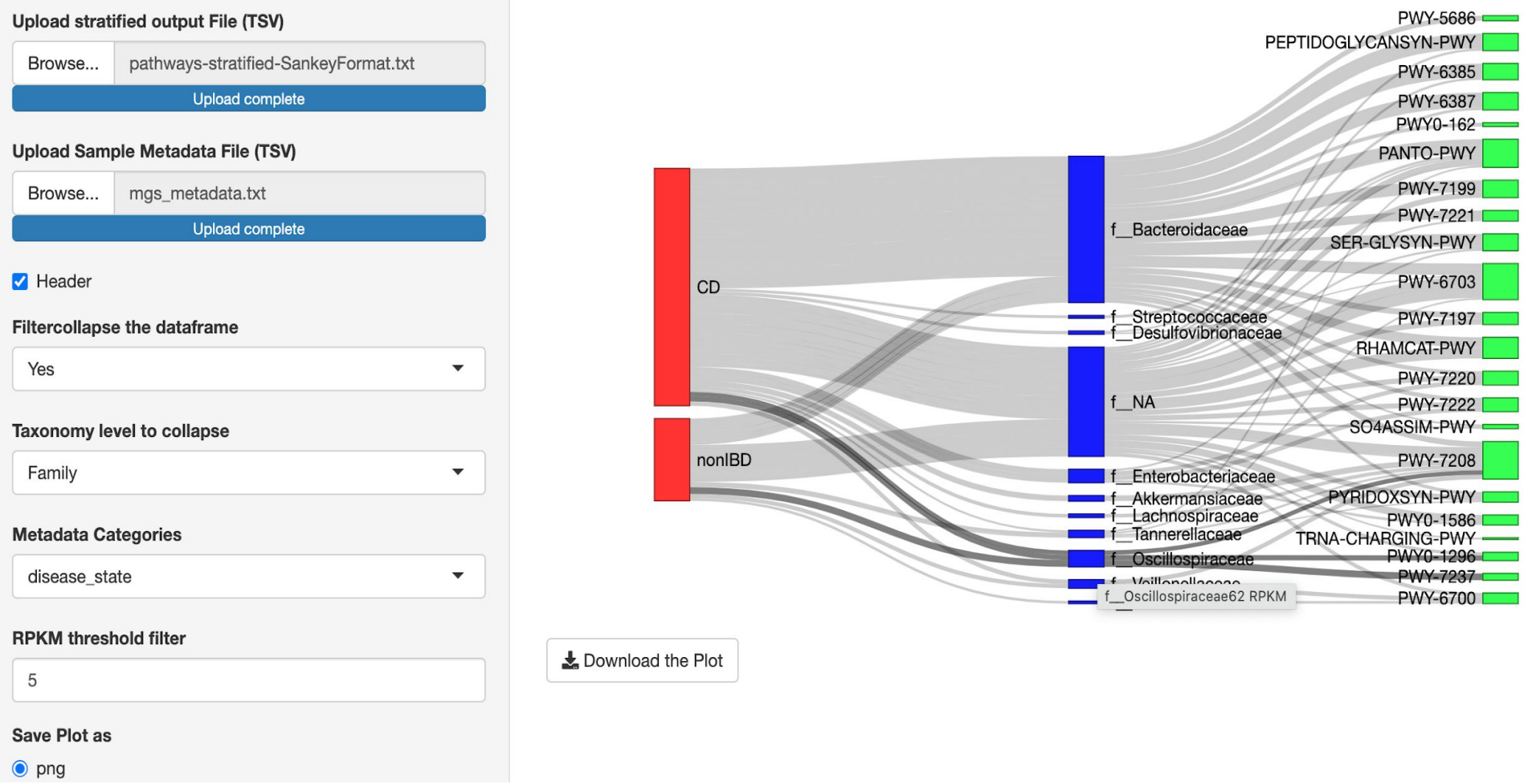
Functional Contributions



Douglas & Langille, 2021

JARRVIS

- Just Another stRatified Rpkm VISualizer



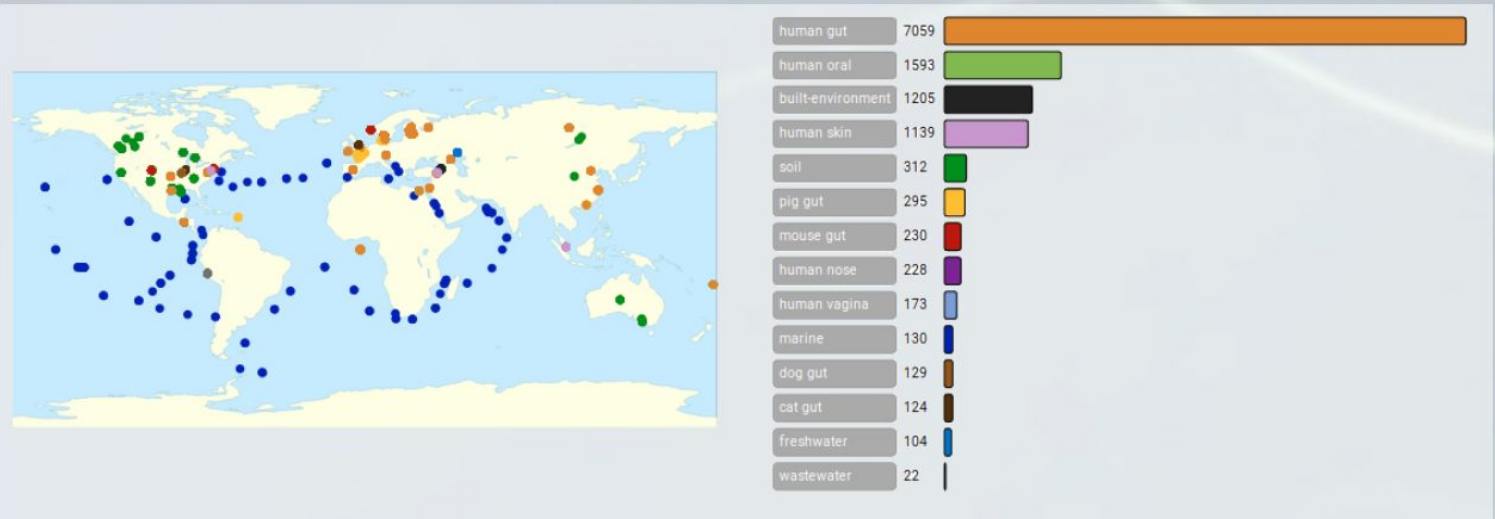
Specialized gene annotation systems/dbs

- Depending on research question, other specialized tools and databases can be used
- Virulence factors
  - VFDB
- Antimicrobial resistance (AMR) genes
  - CARD
- Carbohydrate metabolizing genes
  - CAZy
- Phage and prophage
  - VIBRANT, VirSorter, VirFinder, and MARVEL
- Large microbiome catalogues
  - Based on clustering of previous MGS reads

Global Microbial Gene Catalog v1.0

The **Global Microbial Gene Catalog** is an integrated, consistently-processed, gene catalog of the microbial world, combining metagenomics and high-quality sequenced isolates. A total of 2.3 billion ORFs from 13,174 metagenomes (covering 14 habitats) and the complete ProGenomes2 database were clustered together at 95% nucleotide identity to build a catalog of 302,655,267 unigenes.

Geographical and habitat distribution of samples used to build this catalogue



Community Function Potential

- Important that meta**genomics,** is not meta**transcriptomics**, and not meta**proteomics**
- These annotations suggest the functional **potential** of the community
- The presence of these genes/functions does not mean that they are biologically active (e.g. may not be transcribed)
- DNA may also be from dead cells

We are on a Coffee Break & Networking Session

Workshop Sponsors: