MEDICAL MICROBIOLOGY AND INFECTIOUS DISEASES BIOINFORMATICS WORKSHOP

Presents

INTRODUCTION TO BACTERIAL GENOMICS:
Reference Databases and Taxonomic Classification

INSTRUCTED BY

Jill Rumore, PhD Candidate

Department of Medical Microbiology and Infectious Diseases

University of Manitoba

INFORMATION FOR PARTICIPANTS

All workshops are being recorded and posted to the MMID Bioinformatics Workshop – YouTube

For live Q&A, go to slido.com and use participant code #1888233

2023 MMID Bioinformatics Workshop Schedule

DATE	INSTRUCTOR	TOPIC
March 2	Grace E. Seo	Introduction to the 2023 MMID Bioinformatics Workshop
March 9	Grace E. Seo	Introduction to conda and tool installation
March 16	Grace E. Seo	Introduction to genomics and viral data analysis
March 23	Jill Rumore	Bacterial Genomics
March 30	Jill Rumore	Reference Databases
April 6	Taylor Davedow	Beginner's Guide to Phylogenetic Trees
April 13	Taylor Davedow	Introduction to tree visualization and annotation using ggtree
April 20	-	Bfx workshop: Bring your own dataset!
April 27	-	Bfx workshop: Bring your own dataset!

April 20 and April 27 in-person sessions are open to the public (up to 100 people)!

Work with your colleagues/friends to analyze data together.

SET UP WI-FI (IN-PERSON PARTICIPANTS)

1. ***Connect to UofM-secure (if you are a student or staff)***
   - Use your @myumanitoba.ca or @umanitoba.ca login and password

2. ***Connect to UofM-guest***

To access uofm-guest Wi-Fi:

1. Ensure your wireless card is active and connected to the **uofm-guest** network.
2. Open your web browser (e.g. Google Chrome, Microsoft Edge, Firefox, etc.) and browse to any website. This should redirect you to the **Acceptable Use Agreement** page.
3. Review the Acceptable Use Agreement for the unsecured wireless.
4. Select **I Agree**.

LEARNING OBJECTIVES

1. *What is taxonomic classification?*
2. *What are reference databases?*
3. *Challenges with reference databases*
4. *How to choose classification software*
5. *Perform taxonomic classification on a publically available dataset using Kraken 2*
6. *Visualize results in Pavian (time permitting)*

PUBLICALLY AVAILABLE DATASET

https://doi.org/10.3389/fpubh.2019.00066

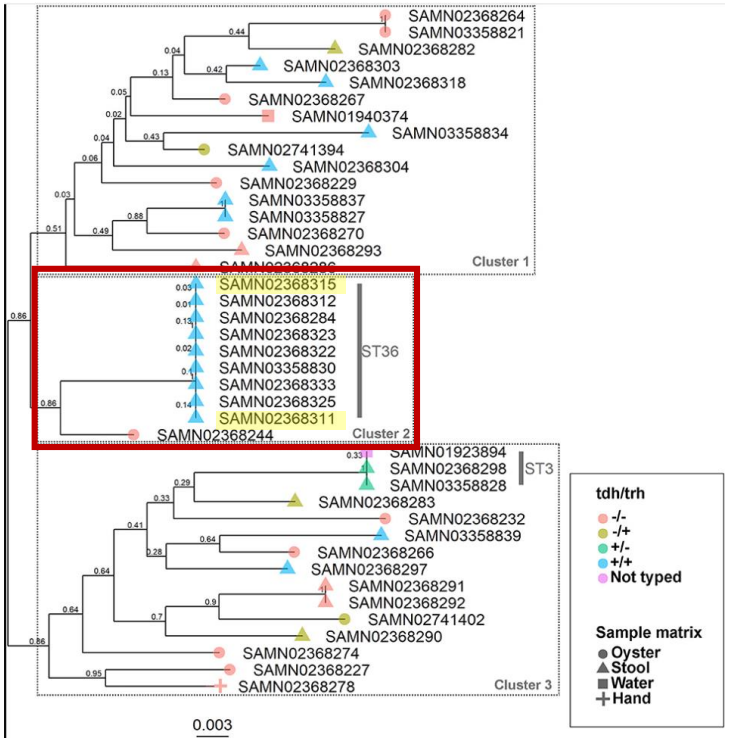
ORIGINAL RESEARCH article

Front. Public Health, 08 May 2019
Sec. Infectious Diseases: Epidemiology and Prevention
Volume 7 - 2019 | https://doi.org/10.3389/fpubh.2019.00066

Clustering of *Vibrio parahaemolyticus* Isolates Using MLST and Whole-Genome Phylogenetics and Protein Motif Fingerprinting

Kelsey J. Jesser[1*], Willy Valdivia-Granda[2], Jessica L. Jones[3] and Rachel T. Noble[1]

[1] Institute of Marine Sciences, University of North Carolina at Chapel Hill, Morehead City, NC, United States
[2] Orion Integrated Biosciences, New Rochelle, NY, United States
[3] Gulf Coast Seafood Laboratory, Division of Seafood Science and Technology, U.S. Food and Drug Administration, Dauphin Island, AL, United States



GETTING STARTED

Please note: these steps should be completed prior to the workshop.

1. Open your terminal and navigate to the conda_workshop directory

```
cd /mnt/c/Users/JRumore/Desktop/conda_workshop
```

2. Make a new directory called Reference_Databases

```
mkdir Reference_Databases
```

3. Open your internet browser and navigate to the MMID Bioinformatics Workshop Github 2023-03-30-Reference-Databases repository (https://github.com/mmid-bioinformatics-workshop/2023-03-30-Reference-Databases) and download the workshop datasets to the Reference_Databases directory.

```
https://drive.google.com/drive/folders/1vVc2KJnlAsy8u2l6VPgKWMEBtmL6zRwE
?usp=share_link
```

4. From the same repository, download the Kraken 2 database to the Reference_Databases directory.

```
https://drive.google.com/drive/folders/1Lzdpl6XW4anl4lNtU5dN2Zuszb44FDJB
?usp=share_link
```

**This will take ~ 20 minutes to download*

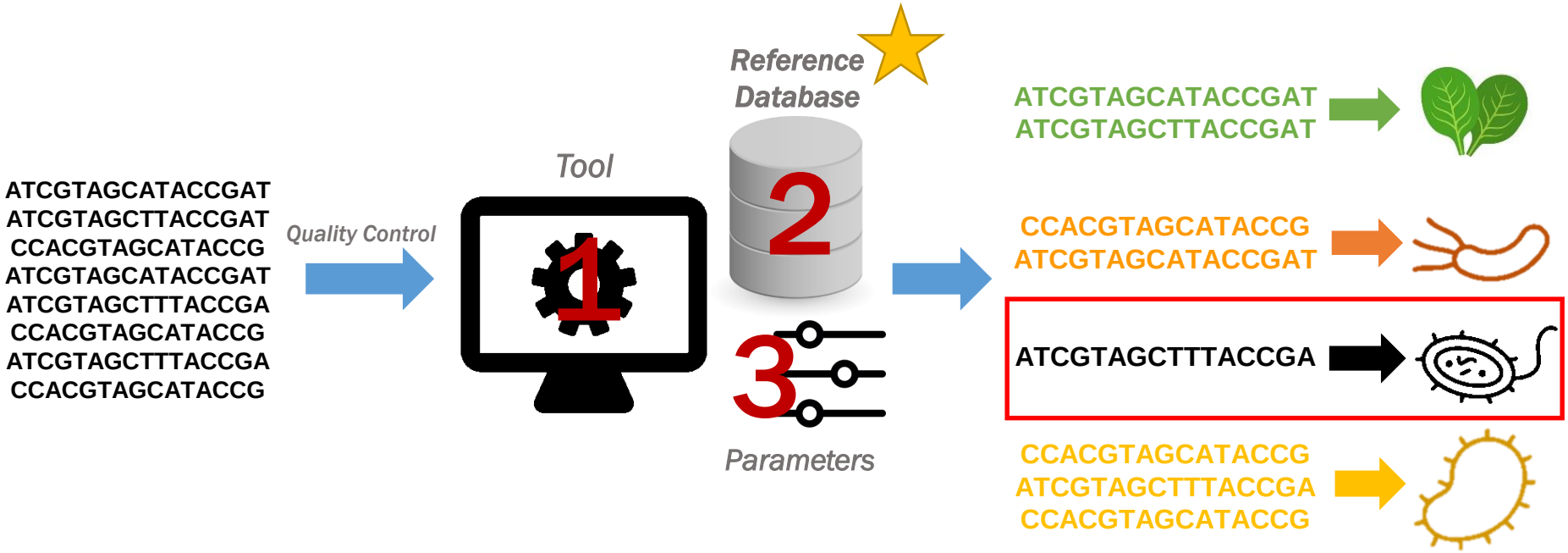
GETTING STARTED

5. Move into the Reference_Databases directory and decompress the Kraken 2 database file (i.e., kraken2_STND-DB-8GB-001.tar.bz2).

```
cd Reference_Databases
tar -xvf kraken2_STND-DB-8GB-001-tar.bz2
```

TAXONOMIC CLASSIFICATION

Read-based classification workflow



RAW SEQUENCE DATA

TAXONOMIC CLASSIFICATION

PATHOGEN ID

REFERENCE DATABASES

- Most classifiers are distributed with pre-compiled reference databases.
- Database content commonly comes from RefSeq Complete Genomes, BLAST nt or nr
- Majority of tools allow user to build custom database
  - Greater control over analysis
  - Can be computationally intensive

https://www.ncbi.nlm.nih.gov/refseq/

CHALLENGES

- Contamination and incompleteness can lead to both false positive and false negative results
  - Considerable amount of contamination in publically available sequence repositories
  - Reads without a reference in the database may be labelled as unknown or imprecisely assigned to the next closest taxon


PLOS ONE

OPEN ACCESS PEER-REVIEWED
RESEARCH ARTICLE

Abundant Human DNA Contamination Identified in Non-Primate Genome Databases

Mark S. Longo, Michael J. O'Neill, Rachel J. O'Neill

Published: February 16, 2011 • https://doi.org/10.1371/journal.pone.0016410


PeerJ

Unexpected cross-species contamination in genome sequencing projects

Research article Bioinformatics Computational Biology Genomics Microbiology

Samier Merchant[1,2], Derrick E. Wood[1,3], Steven L. Salzberg[1,3,4]

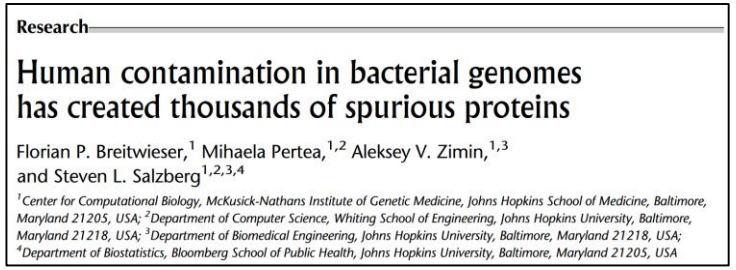
Published November 20, 2014


Commentary | Open Access | Published: 30 March 2015

Large-scale contamination of microbial isolate genomes by Illumina PhiX control

Supratim Mukherjee, Marcel Huntemann, Natalia Ivanova, Nikos C Kyrpides & Amrita Pati

Standards in Genomic Sciences **10**, Article number: 18 (2015) | Cite this article

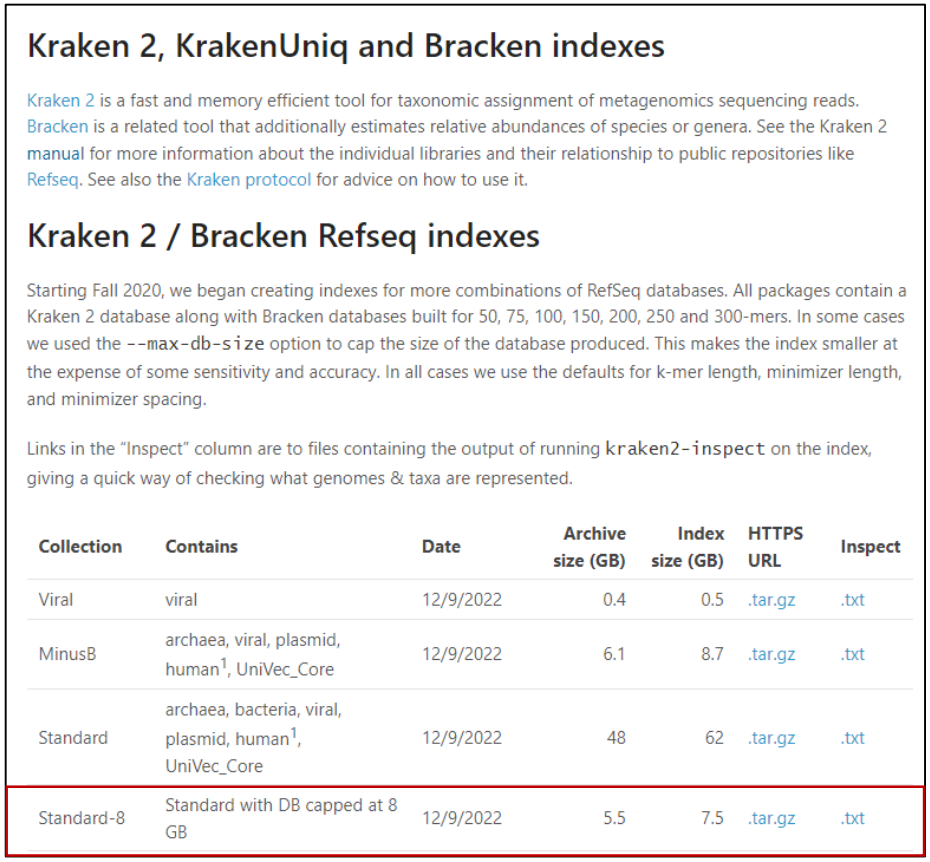

Research

Human contamination in bacterial genomes has created thousands of spurious proteins

Florian P. Breitwieser,[1] Mihaela Pertea,[1,2] Aleksey V. Zimin,[1,3] and Steven L. Salzberg[1,2,3,4]

[1]*Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland 21205, USA;* [2]*Department of Computer Science, Whiting School of Engineering, Johns Hopkins University, Baltimore, Maryland 21218, USA;* [3]*Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland 21218, USA;* [4]*Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland 21205, USA*

PRE-COMPUTED REFERENCE DATABASES



Kraken 2, KrakenUniq and Bracken indexes

Kraken 2 is a fast and memory efficient tool for taxonomic assignment of metagenomics sequencing reads. Bracken is a related tool that additionally estimates relative abundances of species or genera. See the Kraken 2 manual for more information about the individual libraries and their relationship to public repositories like Refseq. See also the Kraken protocol for advice on how to use it.

Kraken 2 / Bracken Refseq indexes

Starting Fall 2020, we began creating indexes for more combinations of RefSeq databases. All packages contain a Kraken 2 database along with Bracken databases built for 50, 75, 100, 150, 200, 250 and 300-mers. In some cases we used the `--max-db-size` option to cap the size of the database produced. This makes the index smaller at the expense of some sensitivity and accuracy. In all cases we use the defaults for k-mer length, minimizer length, and minimizer spacing.

Links in the "Inspect" column are to files containing the output of running `kraken2-inspect` on the index, giving a quick way of checking what genomes & taxa are represented.

Collection	Contains	Date	Archive size (GB)	Index size (GB)	HTTPS URL	Inspect
Viral	viral	12/9/2022	0.4	0.5	.tar.gz	.txt
MinusB	archaea, viral, plasmid, human[1], UniVec_Core	12/9/2022	6.1	8.7	.tar.gz	.txt
Standard	archaea, bacteria, viral, plasmid, human[1], UniVec_Core	12/9/2022	48	62	.tar.gz	.txt
Standard-8	Standard with DB capped at 8 GB	12/9/2022	5.5	7.5	.tar.gz	.txt

https://benlangmead.github.io/aws-indexes/k2

CASE STUDY #1

Incompleteness in the standard reference database can skew results.

Standard databases are not one size fits all:

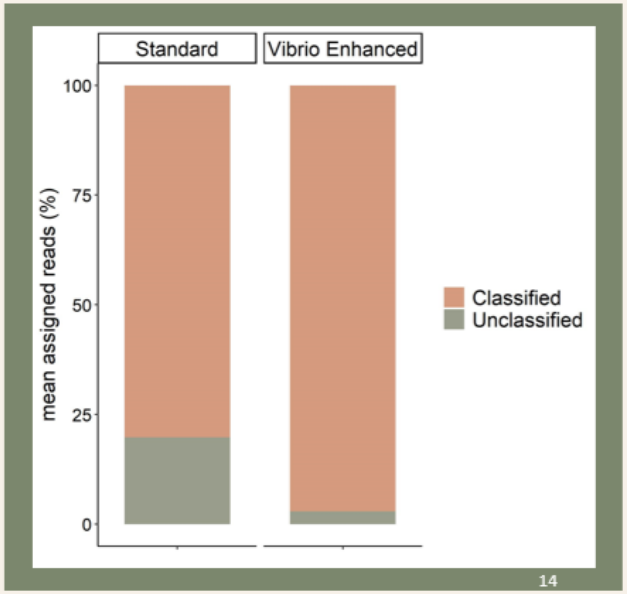
genus representation matters

Kraken2 Standard database

- *Vibrio* species represents **1.35%**
- Missing species: albensis, aestuarianus, brasiliensis, ordalii and shilonii

***Vibrio* Enhanced Kraken2 database**

- *Vibrio* species represents **2.01%**
- Improved read classification
- 54/55 below 5% unclassified
- 12/13 unclassified samples were classified
- One sample remained unclassified

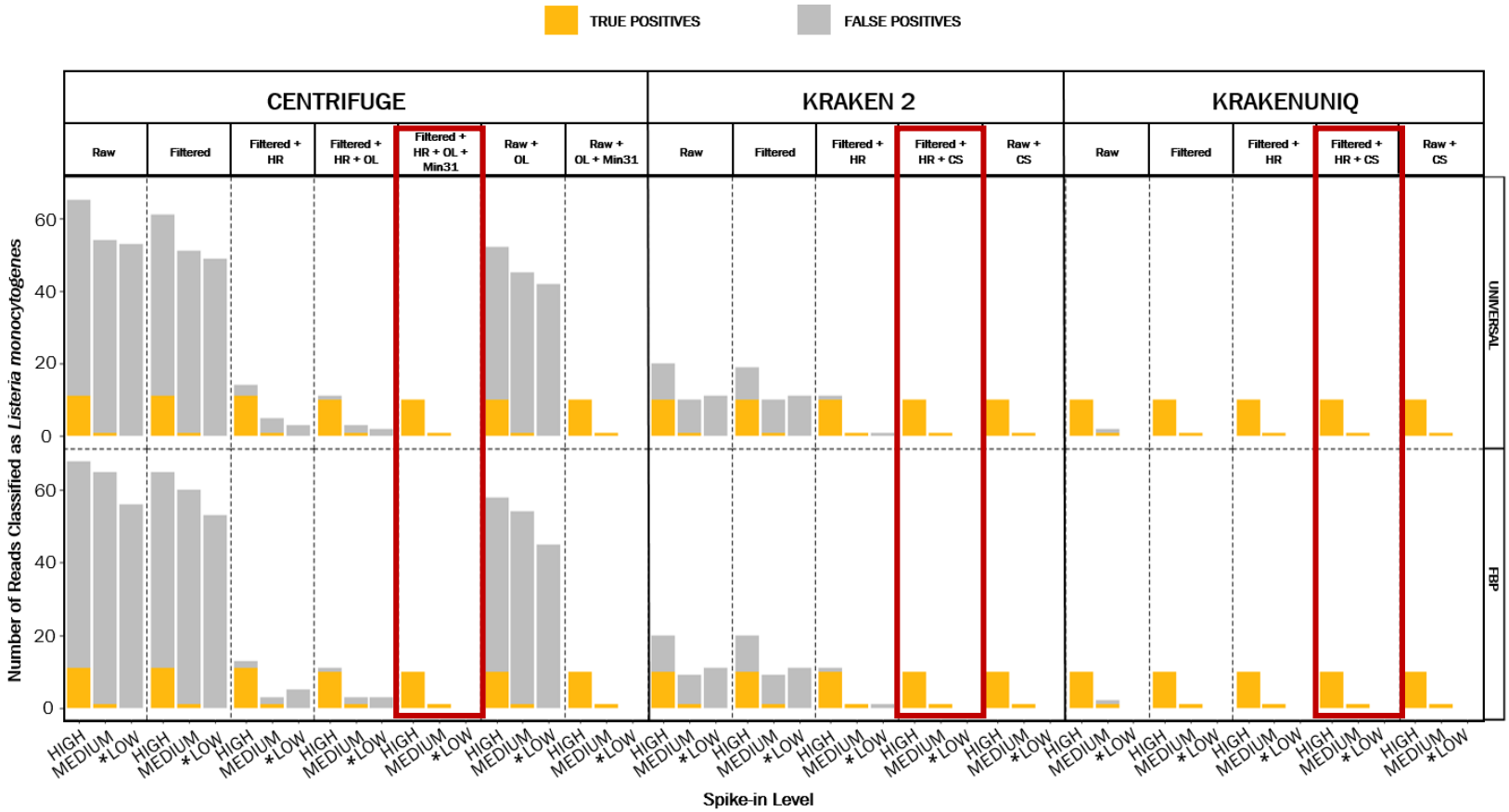


Database creation and collaboration: Jill Rumore

Slide Credit: Taylor Davedow, 2022

CASE STUDY #2

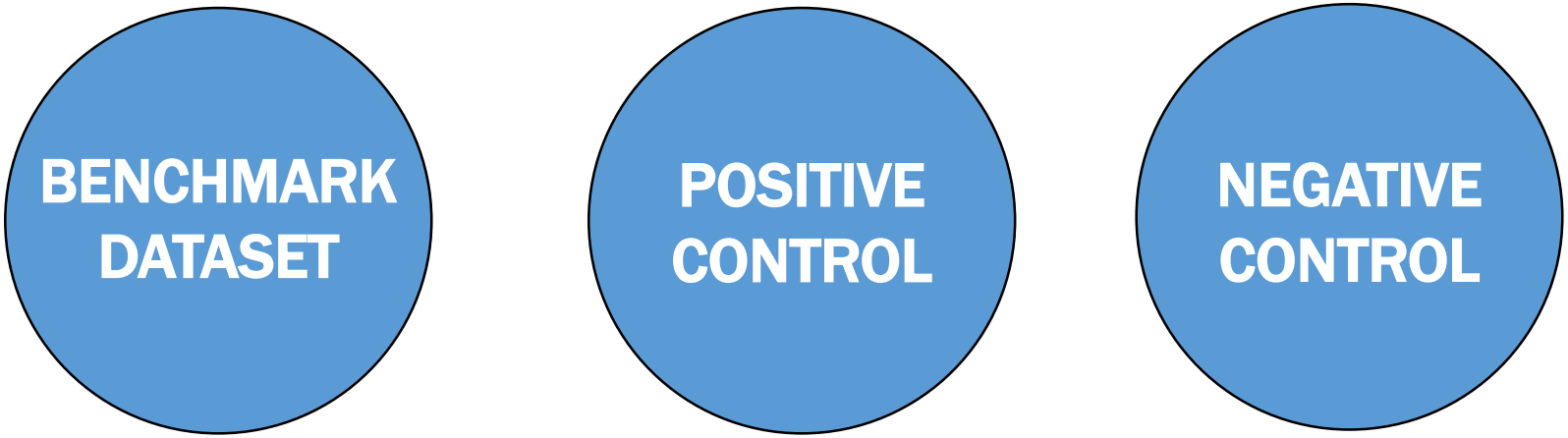
A standard (universal) reference database is not always necessary for accurate and reliable results.



Rumore et al., 2023; *BMC Genomics – Submitted*

TAKE HOME MESSAGE

ALWAYS EVALUATE AND VALIDATE YOUR REFERENCE DATABASE!



CUSTOM REFERENCE DATABASES

Considerations for building your own reference database

1. Use complete or quality-controlled genomes
    I. Screen reference genomes using checkM for completeness and contamination

2. Mask low complexity sequences (e.g. ACACACACACACACACACACACACAC)
    I. Automated for some classification software when downloading from NCBI (i.e., Kraken 2, KrakenUniq)
    II. Dustmasker – included in BLAST suite of tools (https://www.liebertpub.com/doi/10.1089/cmb.2006.13.1028)

3. Filter out contigs that are less than 1000 bp when using draft genomes
    I. Study found that majority of contaminated contigs are < 1000 bp (https://pubmed.ncbi.nlm.nih.gov/31064768/)
    II. Use `seqkit seq -m 1000 reference.fasta` (https://anaconda.org/bioconda/seqkit)

4. Include the human genome
    I. #1 contaminant in the lab!

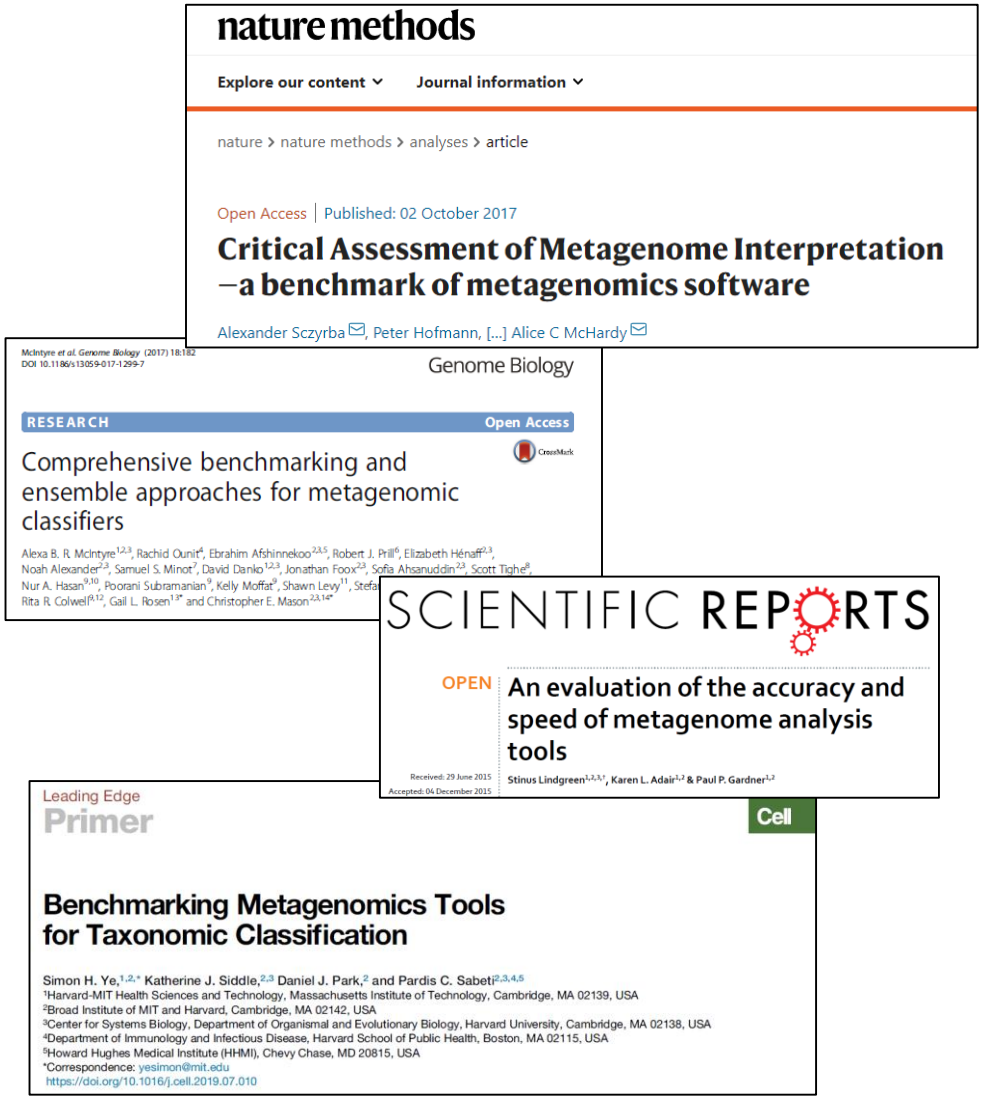
5. Include the contaminant databases UniVec and EMVEC
    I. Automated download for some classification software (i.e., Kraken 2, KrakenUniq)
    II. UniVec (https://ftp.ncbi.nlm.nih.gov/pub/UniVec/)
    III. EMVEC (https://ftp.ebi.ac.uk/pub/databases/emvec/)

TOOL SELECTION CRITERIA

AVAILABILITY

USABILITY

ADOPTION



Ye et al., 2019; Cell
Lindgreen et al., 2016; Scientific Reports

CLASSIFICATION SOFTWARE

Three Versions:

Kraken – **No longer supported**

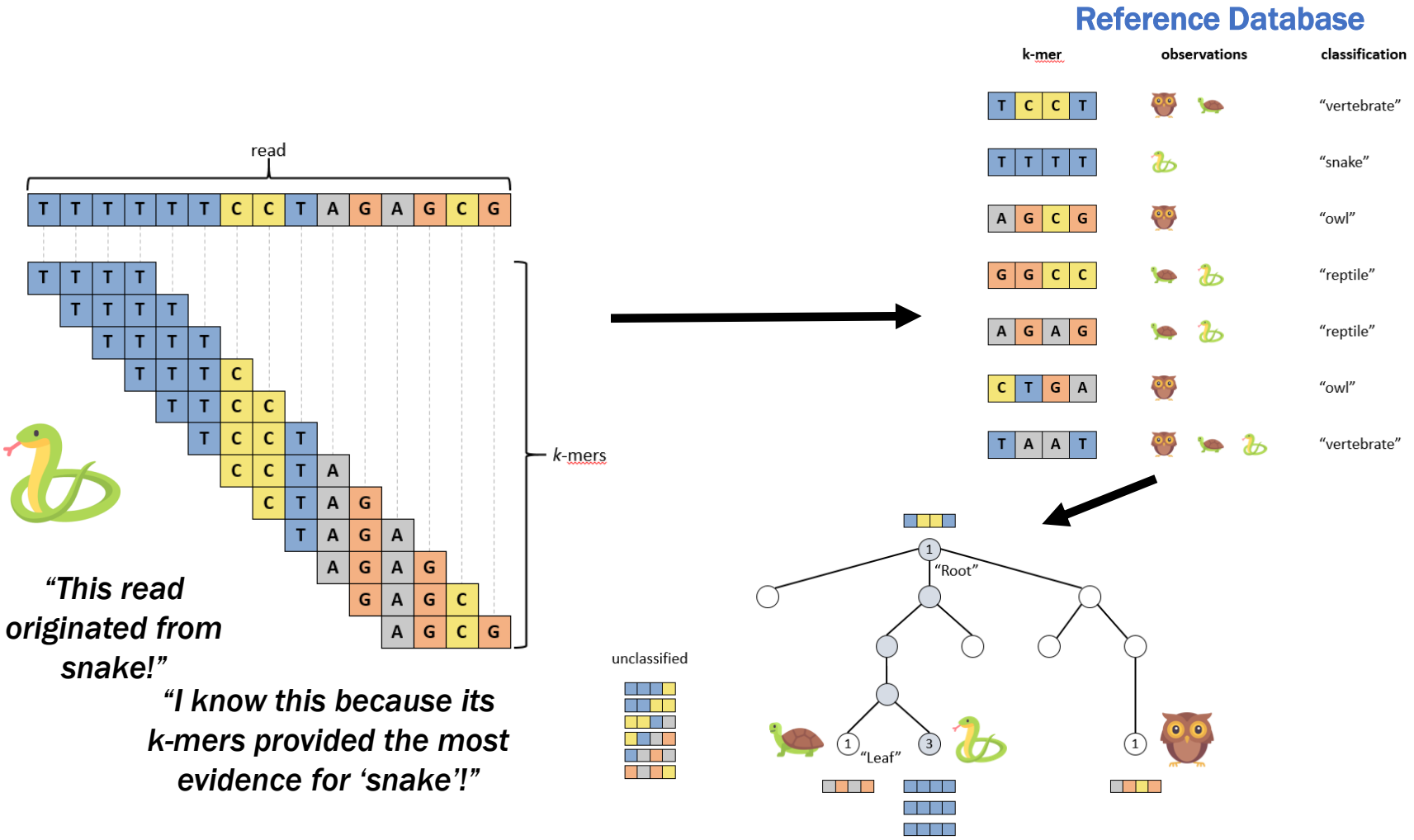
Kraken2 (https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1891-0)

- -More memory efficient (~85% less memory than KrakenUniq)
- -Uses smaller databases (runs ~ 5X faster than KrakenUniq)
  - -More false-positive classifications (though minimal) are possible
- -Not compatible with original Kraken databases
- -New Feature = unique *k*-mers (Kraken2Uniq)
  - -use the `--report-minimizer-data` flag to force Kraken 2 to provide unique *k*-mer counts

KrakenUniq (https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1568-0)

- -Compatible with original Kraken databases
- -Memory intensive
- -Large reference databases
- -Reports unique *k*-mers

k-mer BASED APPROACH

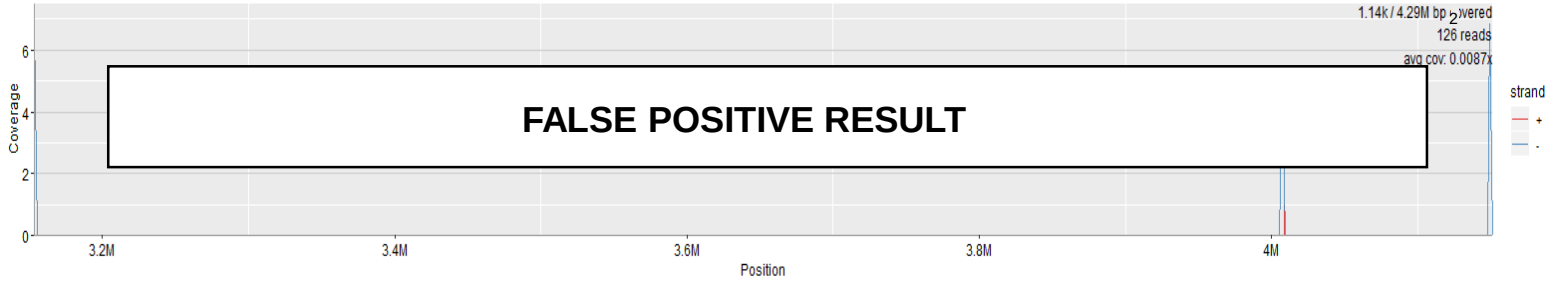


Slide Credit: Eric Marinier, 2018

Unique *k*-mers

Reads = 15 Unique *k*-mers = 1570

Reads = 122 Unique *k*-mers = 126



Unique *k*-mers = proxy for genome coverage

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1568-0

KRAKEN 2 DATABASES

A Kraken 2 database is a directory containing at least 3 files:

hash.k2d: Contains the minimizer to taxon mappings
opts.k2d: Contains information about the options used to build the database
taxo.k2d: Contains taxonomy information used to build the database

Check the contents of the database

```
kraken2-inspect --db ./kraken2_STND-DB_8GB/ > k2-inspect.txt
```

100	1.4E+09	521261	R	1	root
99.18	1.39E+09	261134	R1	131567	cellular organisms
89.58	1.25E+09	2647675	D	2	Bacteria
43.79	6.13E+08	2093873	P	1224	Proteobacteria
21.37	2.99E+08	1246636	C	1236	Gammaproteobacteria
5.17	72488581	1195303	O	91347	Enterobacterales
2.49	34915053	2879049	F	543	Enterobacteriaceae
0.46	6391862	122031	F1	2890311	Klebsiella/Raoultella group
0.38	5310235	1569892	G	570	Klebsiella
0.08	1154255	1109728	S	573	Klebsiella pneumoniae
0	31093	26989	S1	72407	Klebsiella pneumoniae subsp. pneumoniae
0	1369	1369	S2	1328324	Klebsiella pneumoniae subsp. pneumoniae KPNIH27
0	1311	1311	S2	272620	Klebsiella pneumoniae subsp. pneumoniae MGH 78578
0	676	676	S2	1123862	Klebsiella pneumoniae subsp. pneumoniae Kp13
0	365	365	S2	1193292	Klebsiella pneumoniae subsp. pneumoniae 1084
0	119	119	S2	1392499	Klebsiella pneumoniae subsp. pneumoniae 1158

KRAKEN 2 OUTPUT

Two Main Output Files

Read Classification (Standard Output)

Classified (C)/ Unclassified (U)	Sequence ID	taxID	Sequence Length (bp)	k-mer mapping
C	SRR1815541.34	670	100|100	670:2 0:5 670:11 0:29 717610:1 0:18 |:| 0:20 717610:3 0:41 717610:1 0:1
C	SRR1815541.41	670	96|96	670:5 0:8 670:2 0:26 670:4 0:17 |:| 0:17 670:4 0:26 670:2 0:8 670:5
C	SRR1815541.44	670	100|100	670:4 0:54 670:8 |:| 670:4 0:59 670:3

Report

% Reads	reads	taxReads	minimizers	unique *k*-mers	rank	taxID	taxName
7.79	321458	321458	0	0	U	0	unclassified
92.21	3807168	3452	20282678	263741	R	1	root
92.13	3803710	114	20253832	263741	R1	131567	cellular organisms
92.12	3803448	4900	20248643	263427	D	2	Bacteria
92	3798391	2343	20127817	262002	P	1224	Proteobacteria
91.94	3795979	28736	20077116	260635	C	1236	Gammaproteobacteria
91.23	3766666	0	19633653	253556	O	135623	Vibrionales
91.23	3766666	12935	19633653	253556	F	641	Vibrionaceae
90.92	3753652	446430	19438434	250091	G	662	Vibrio
80.05	3304914	248124	14994313	188884	G1	717610	Vibrio harveyi group
73.96	3053658	3052752	13604428	170278	S	670	Vibrio parahaemolyticus
0.01	460	460	860	142	S1	1211705	Vibrio parahaemolyticus BB22OP
0.01	262	262	381	129	S1	1429044	Vibrio parahaemolyticus UCM-V493

https://github.com/DerrickWood/kraken2

DEMONSTRATION

STEP BY STEP GUIDE

1. Move into the Reference_Databases directory and decompress the Datasets folder downloaded from the MMID Bioinformatics Github repository.

```
unzip Datasets-20230328T145901Z-001.zip
```

2. Move into the Downsampled_HR_fastq directory and decompress the host filtered fastq files.

```
cd ./Datasets/Downsampled_HR_fastq
gunzip *.gz
```

3. Return to the Reference_Databases directory

```
cd ../../
```

4. Make a new directory called kraken2_output

```
mkdir kraken2_output
```

STEP BY STEP GUIDE

5. Activate the conda environment containing the Kraken 2 package and review the contents of the environment to ensure the tool is installed.

```
conda activate conda_workshop
conda list
```

6. Review the Kraken 2 man page

```
kraken2 --help
```

7. Run Kraken 2 from the Reference_Databases directory using the test dataset SAMN02368311

```
kraken2 --db ./kraken2_STND-DB-8GB/ --threads 2 --report
./kraken2_output/SAMN02368311-K2reportfile.tsv --report-minimizer-
data --paired ./Datasets/Downsampled_HR_fastq/SAMN02368311_R1.fastq
./Datasets/Downsampled_HR_fastq/SAMN02368311_R2.fastq >
./kraken2_output/SAMN02368311-K2readclassification.tsv
```

PAVIAN

Bioinformatics. 2020 Feb 15;36(4):1303-1304. doi: 10.1093/bioinformatics/btz715.

Pavian: interactive analysis of metagenomics data for microbiome studies and pathogen identification

Florian P Breitwieser [1], Steven L Salzberg [2]

Affiliations + expand

PMID: 31553437 PMCID: PMC8215911 DOI: 10.1093/bioinformatics/btz715

Free PMC article

Install Pavian in R

https://github.com/fbreitwieser/pavian

DEMONSTRATION

STEP BY STEP GUIDE

1. Open the Pavian Shiny App.

https://fbreitwieser.shinyapps.io/pavian/

2. Upload the pre-computed Kraken 2 reports.

Click “Browse”
Navigate to Kraken 2_Reports folder containing the *report.tsv files

3. Once the files are uploaded, click “Sample” to visualize the Sankey diagram.

HELPFUL RESOURCES

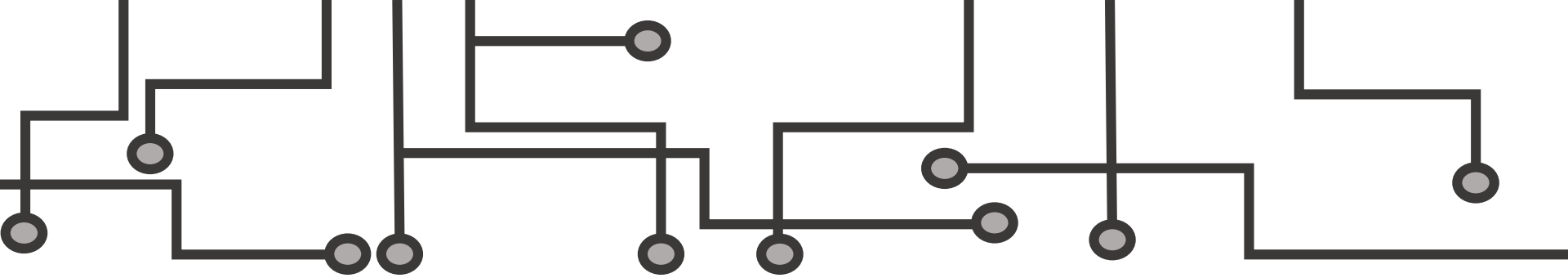
Kraken 2 wiki
https://github.com/DerrickWood/kraken2/wiki/Manual

KrakenUniq Wiki
https://github.com/fbreitwieser/krakenuniq/blob/master/README.md

Step-by-Step Protocol for classification, quantification and visualization
https://www.nature.com/articles/s41596-022-00738-y

How to choose a classification software
http://ccb.jhu.edu/software/choosing-a-metagenomics-classifier/

Kraken Tools
https://github.com/jenniferlu717/KrakenTools



THANK YOU FOR ATTENDING!

Please make sure to fill out the Exit Survey at https://docs.google.com/forms/d/e/1FAIpQLSem_XeuoxBm7E-TLN5E6Vfy0ZVZyBF08AoSRyZaSu_hXfaaQ/viewform?usp=sf_link

We value your feedback!

More questions? Please email us at mmid.bioinformatics.workshop@gmail.com or post them to the workshop slack channel