MEDICAL MICROBIOLOGY AND INFECTIOUS DISEASES BIOINFORMATICS WORKSHOP

Presents

Beginner's Guide to Phylogenetic Trees

INSTRUCTED BY
Taylor Davedow, PhD Student
Department of Medical Microbiology and Infectious Diseases
University of Manitoba

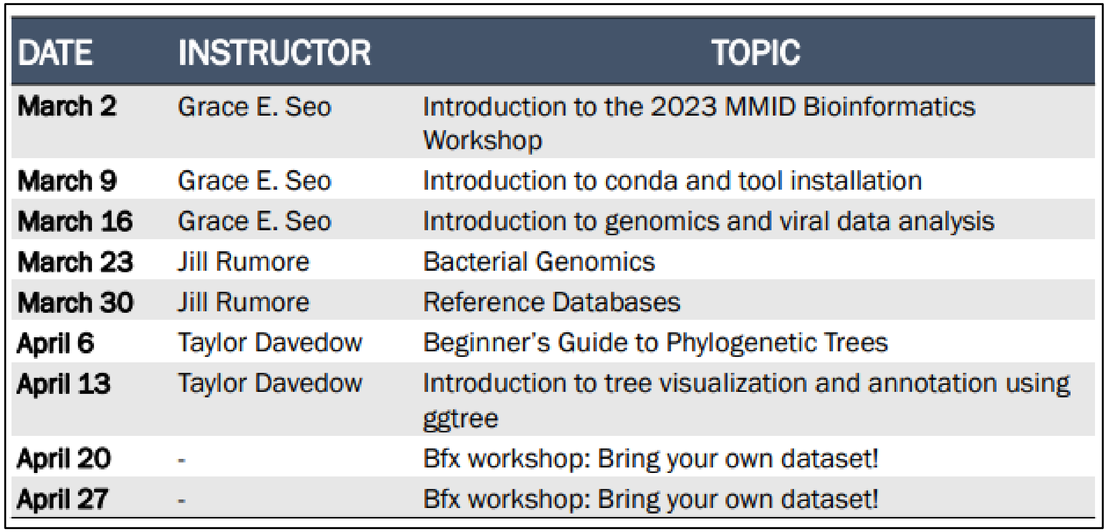
Educational series supported by the Department of MMID to promote knowledge translation and dissemination among students and researchers

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 #3807206

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!

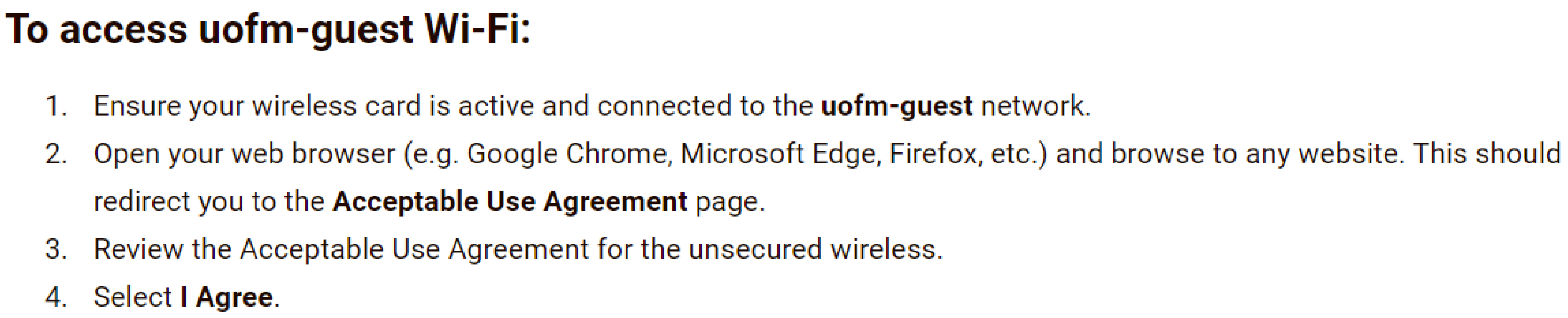
April 20 and April 27 in-person sessions are open to the public (up to 100 people)!

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. *Use a publically available dataset to:*
   I. *Build a phylogenetic tree*

2. Provide a quick view of the tree output

DISCLAIMER

To provide a basic working instruction, all tools will be run with default settings. HOWEVER, careful consideration of analysis parameters in the context of the research question should be taken into account when analyzing your own datasets, as default parameters do not always provide the most optimal result.

GETTING STARTED…

1. Open up the terminal and navigate to your workshop directory

```
cd /mnt/c/Users/Username/Desktop/*insert_dir_name*
```

2. Make a directory called Phylo

```
mkdir Phylo
```

3. List the contents of the directory to confirm the new directory has been created

```
ls
```

GETTING STARTED...

4. Download the multiple sequence alignment fasta file to the Phylo directory

```
https://drive.google.com/file/d/1AR9iopL--g3sf9Uvu83BLNRJhAKxBMz3/view
```

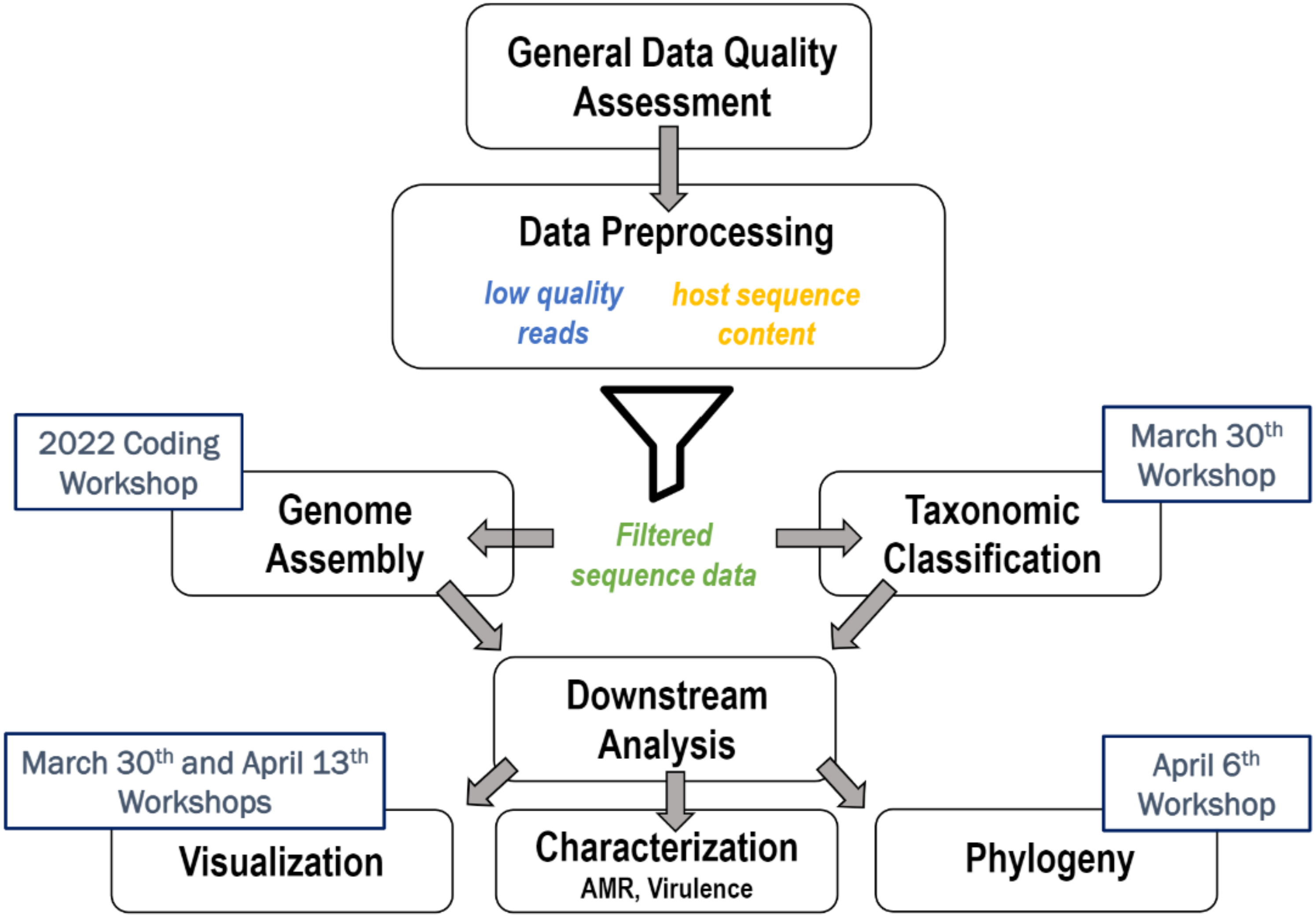
5. Unzip the msa.fasta.zip file

```
unzip msa.fasta.gz
```

6. List the contents of the directory to confirm the file is ready

```
ls
```

BASIC WORKFLOW



General Data Quality Assessment

Data Preprocessing

low quality reads

host sequence content

Filtered sequence data

2022 Coding Workshop

Genome Assembly

March 30th Workshop

Taxonomic Classification

Downstream Analysis

March 30th and April 13th Workshops

Visualization

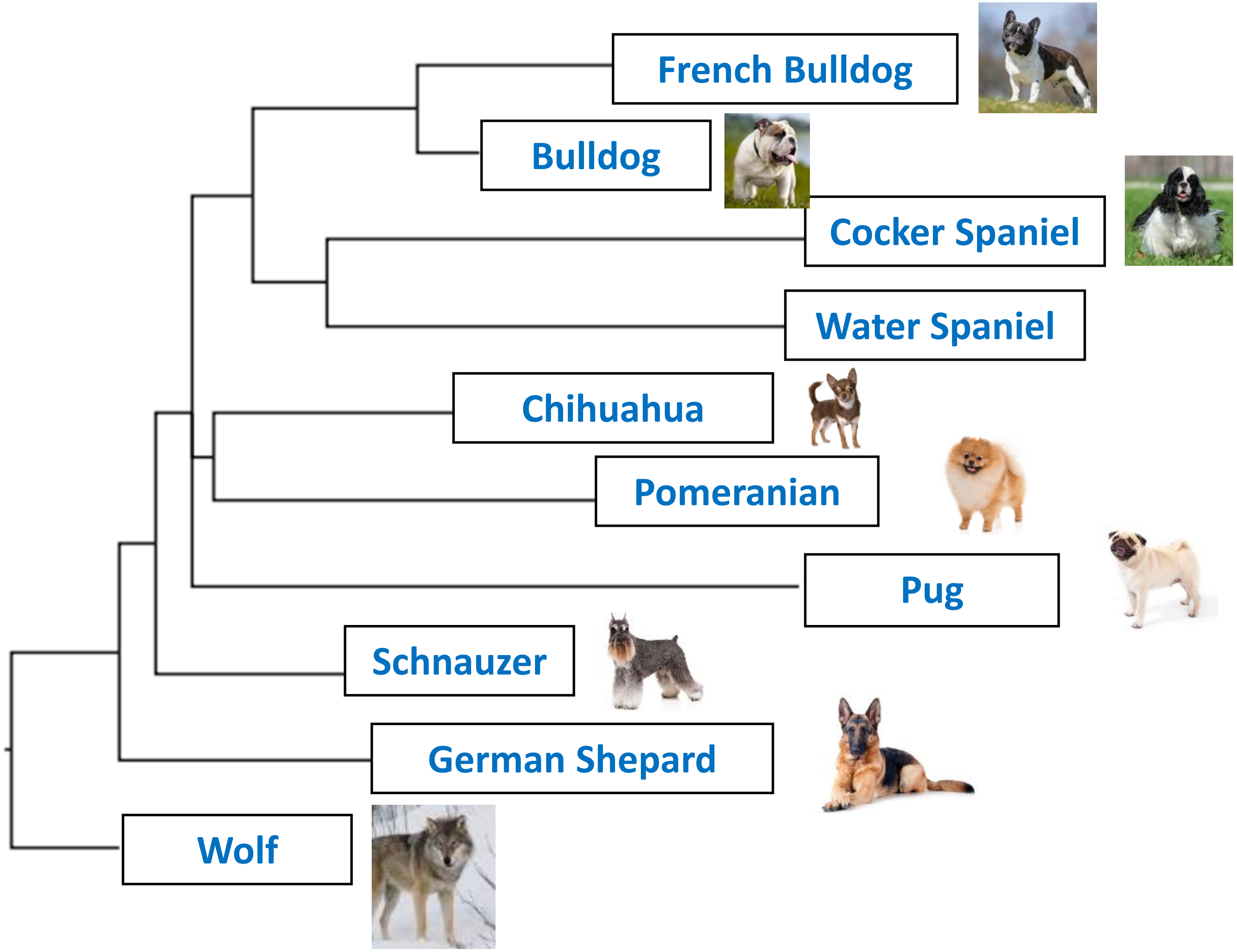
Characterization

AMR, Virulence

April 6th Workshop

Phylogeny

PHYLOGENY



French Bulldog

Bulldog

Cocker Spaniel

Water Spaniel

Chihuahua

Pomeranian

Pug

Schnauzer

German Shepard

Wolf

akc.org, wiki

TYPES OF TREES

Character-based method

Taxa	Characters								
Species A	A	T	G	C	T	A	T	T	C
Species B	A	--	T	C	G	C	T	A	G
Species C	T	--	C	A	C	T	A	G	A

Ex) Maximum-likelihood, Maximum parsimony

Distance-based method

	A	B	C
Species A	----	0.20	0.50
Species B	0.23	----	0.40
Species C	0.87	0.59	----

Ex) neighbor-joining, UPGMA

TREE TOOLS

Name	Description	Methods
IQ-TREE	**An efficient phylogenomic software**	**Maximum likelihood**
MashTree	Rapid comparison of WGS (does not infer phylogeny)	
PhyML	Fast and accurate estimation of phylogenies	Maximum likelihood
QuickTree	Tree construction optimized for efficiency	Neighbor-joining
BEAST	Bayesian Evolutionary Analysis Sampling Trees	Bayesian inference, relaxed molecular clock, demographic history
ClustalW	Progressive multiple sequence alignment	Distance matrix/nearest neighbor
MEGA	Molecular Evolutionary Genetics Analysis	Distance, Parsimony and Maximum Composite Likelihood Methods

DEMONSTRATION

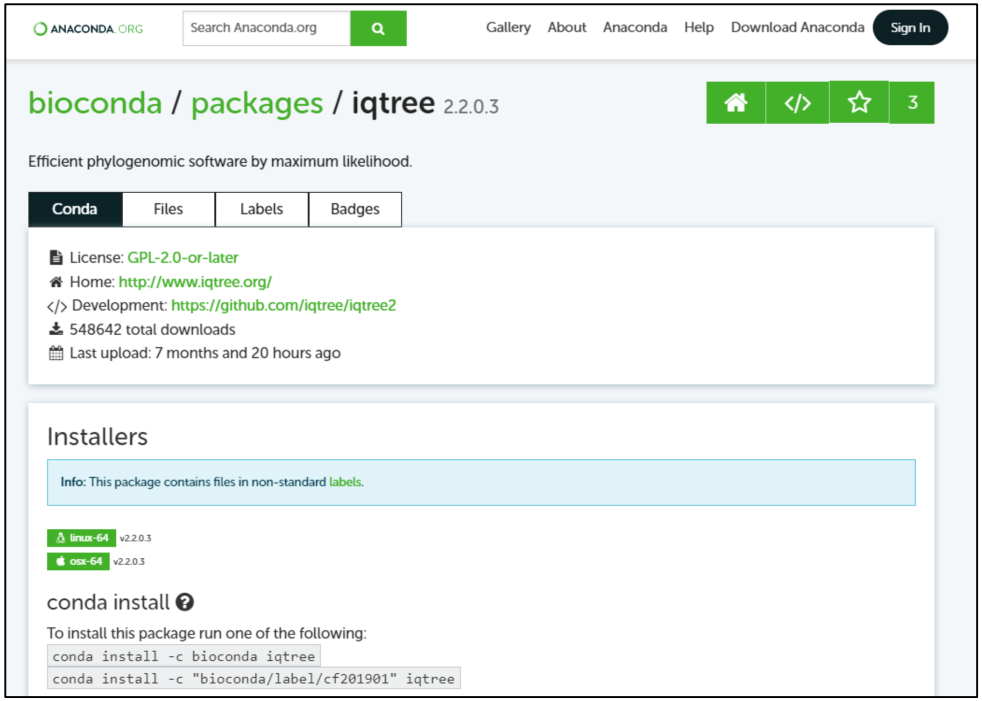
INSTALL PACKAGES

1. **Install iqtree in your workshop environment:**

```
conda install -c bioconda iqtree
```

2. **Check the package was installed:**

```
conda list
```



bioconda / packages / iqtree 2.2.0.3

Efficient phylogenomic software by maximum likelihood.

License: GPL-2.0-or-later
Home: http://www.iqtree.org/
Development: https://github.com/iqtree/iqtree2
548642 total downloads
Last upload: 7 months and 20 hours ago

Installers

Info: This package contains files in non-standard labels.

linux-64 v2.2.0.3
osx-64 v2.2.0.3

conda install

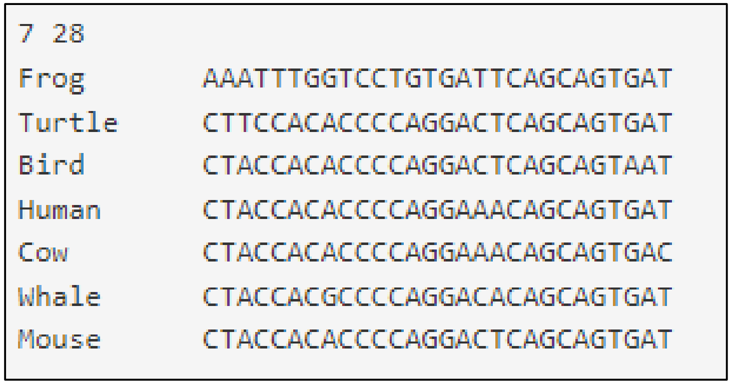
To install this package run one of the following:

```
conda install -c bioconda iqtree
conda install -c "bioconda/label/cf201901" iqtree
```

Multiple Sequence Alignment

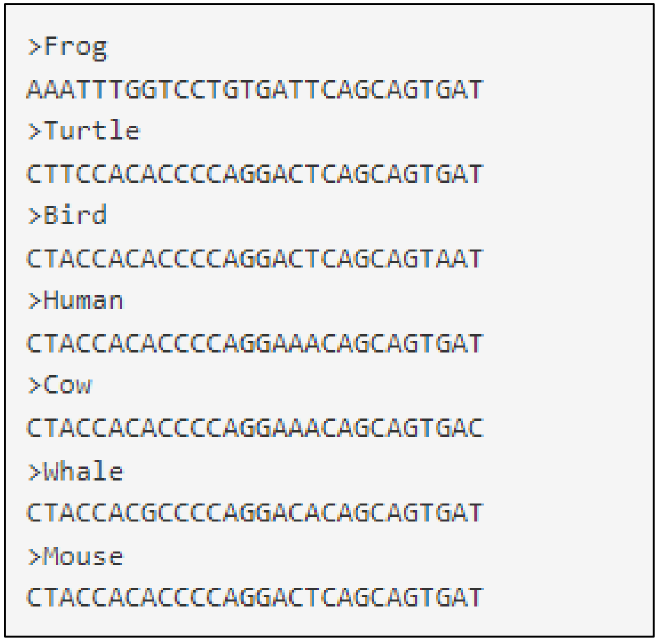
Alignment of biological sequences (protein or nucleic acid) of similar length
Used to infer homology and evolutionary relationships

PHYLIP format



```
7 28
Frog        AAATTTGGTCCTGTGATTCAGCAGTGAT
Turtle      CTTCCACACCCCAGGACTCAGCAGTGAT
Bird        CTACCACACCCCAGGACTCAGCAGTAAT
Human       CTACCACACCCCAGGAAACAGCAGTGAT
Cow         CTACCACACCCCAGGAAACAGCAGTGAC
Whale       CTACCACGCCCCAGGACACAGCAGTGAT
Mouse       CTACCACACCCCAGGACTCAGCAGTGAT
```

FASTA format



```
>Frog
AAATTTGGTCCTGTGATTCAGCAGTGAT
>Turtle
CTTCCACACCCCAGGACTCAGCAGTGAT
>Bird
CTACCACACCCCAGGACTCAGCAGTAAT
>Human
CTACCACACCCCAGGAAACAGCAGTGAT
>Cow
CTACCACACCCCAGGAAACAGCAGTGAC
>Whale
CTACCACGCCCCAGGACACAGCAGTGAT
>Mouse
CTACCACACCCCAGGACTCAGCAGTGAT
```

MSA RESOURCES

1. Clustal Omega

   https://github.com/GSLBiotech/clustal-omega

2. MAFFT

   https://github.com/GSLBiotech/mafft

3. Kalign

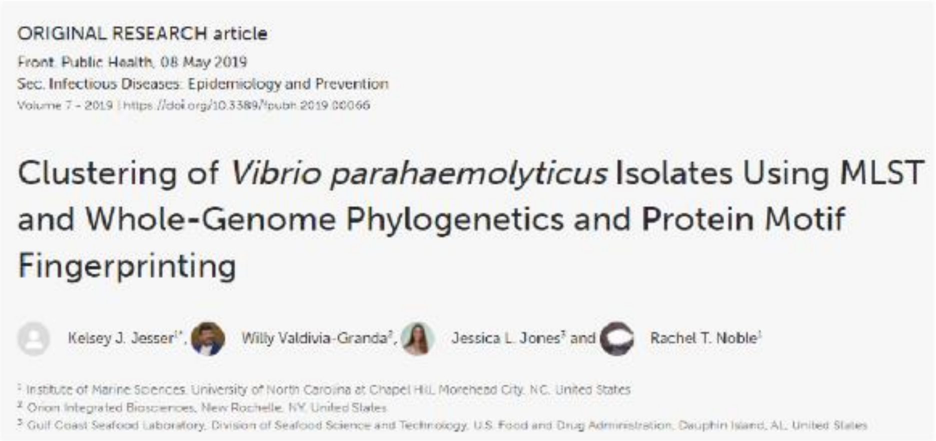
   https://github.com/TimoLassmann/kalign

4. GTDB-TK (Genome Taxonomy Database Toolkit)

   https://github.com/Ecogenomics/GTDBTk

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]

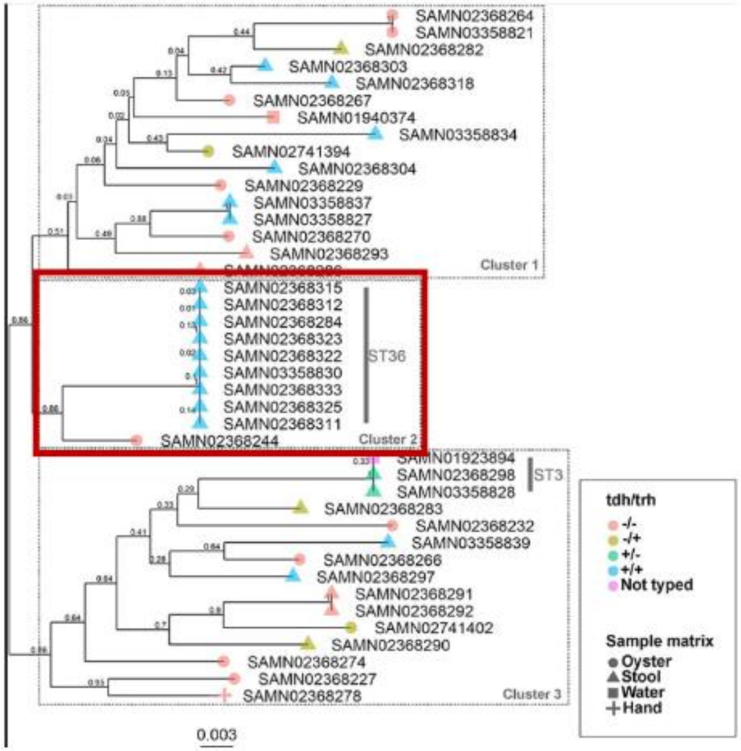
[1] Institute of Marine Sciences, University of North Carolina at Chapel Hill, Morehead City, NC, United States

[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

Uploaded to the MMID Bioinformatics GitHub Repository https://github.com/mmid-bioinformatics-workshop

Data was sequenced using the Illumina HiSeq 2000



DEMONSTRATION

STEP BY STEP GUIDE

1. Make a directory in Phylo called msa

```
mkdir msa
```

2. Move msa file into msa directory

```
mv msa.fasta ./msa
```

3. Verify file was moved

```
cd msa
ls
```

4. Activate conda environment that contains the iqtree package

```
source activate iqtree
```

STEP BY STEP GUIDE

5. Return to Phylo directory

```
cd ..
```

6. Make a new directory called iqtree

```
mkdir iqtree
```

7. Run IQ-TREE

```
cd iqtree
iqtree -s ../msa/msa.fasta
```

****Waffles users only****

```
sbatch -c 1 --mem 48G -p NMLResearch --wrap="iqtree -s
./../msa/msa.fasta"
```

** IF WORKING ON THE CLUSTER (AKA WAFFLES) PLEASE USE THE SLURM WORKLOAD MANAGER WHEN SUBMITTING JOBS**

Detailed instructions can be found here:
https://github.com/MMID-coding-workshop/2022-01-19-Introduction-to-CONDA/blob/main/MMID_Coding_Workshop-IntroToConda_2022-01-19-Supplemental.pdf

IQ-TREE OUTPUT

1. **msa.fasta.iqtree:** the main report file. Will show computational results and contextual representation of final tree.

2. **msa .fasta.treefile:** the ML tree in NEWICK format, which can be visualized by any supported tree viewer program

3. **msa.fasta.log:** Can refer to log file of the entire run to look for errors

DEMONSTRATION

STEP BY STEP GUIDE

1. Check if output for iqtree was generated

```
ls
```

2. Move to msa directory

```
cd ../msa
ls
```

3. Move tree file to iqtree directory and move into that directory

```
mv *.treefile ../iqtree
cd ../iqtree
ls
```

INTERACTIVE TREE VIEWER

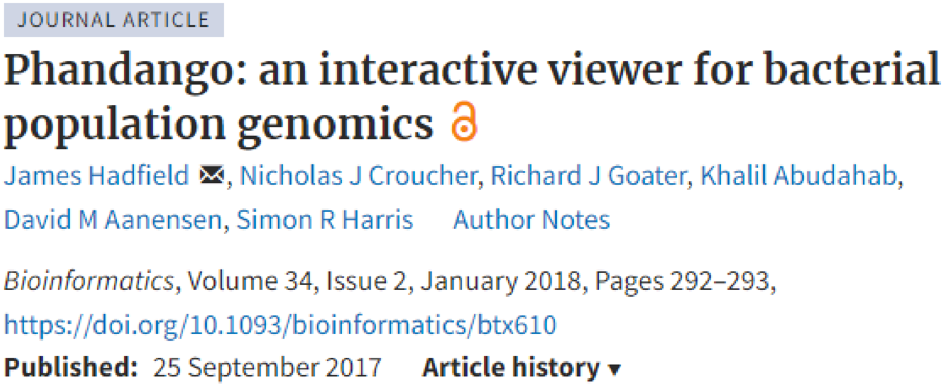
1. Open up web browse and go to phandango: http://jameshadfield.github.io/phandango/#/

2. Open up file explorer and go to the iqtree directory

3. Drag and drop msa.fasta.treefile into phandango

Interactive visualization of genome phylogenies

phandango



JOURNAL ARTICLE

Phandango: an interactive viewer for bacterial population genomics

James Hadfield, Nicholas J Croucher, Richard J Goater, Khalil Abudahab, David M Aanensen, Simon R Harris Author Notes

Bioinformatics, Volume 34, Issue 2, January 2018, Pages 292–293, https://doi.org/10.1093/bioinformatics/btx610

Published: 25 September 2017 **Article history** ▾

DEMONSTRATION

STEP BY STEP GUIDE

1. In the iqtree directory, make a copy of the tree file while changing the extension to specify the file input requirements of phandango

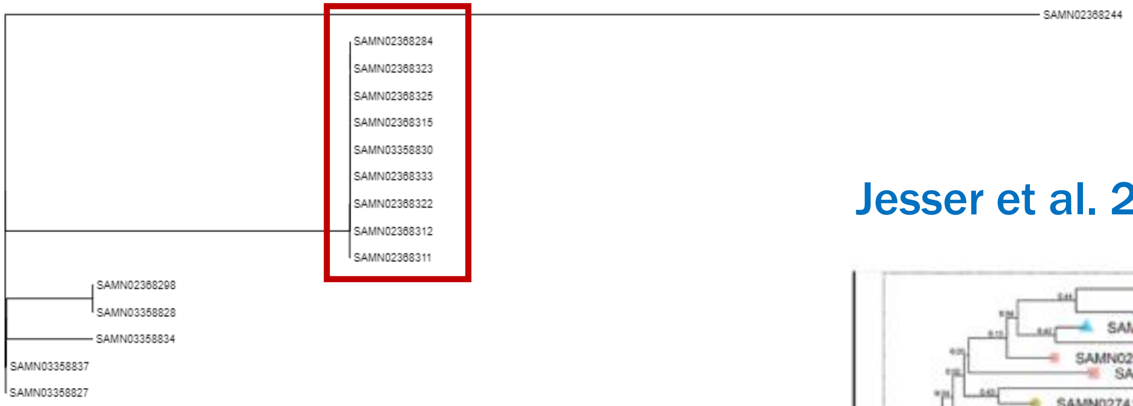
```
cp msa.fasta.treefile msa.fasta.tree (or .tre)
ls
```

2. Drag and drop the new tree file (.nwk) into phandango

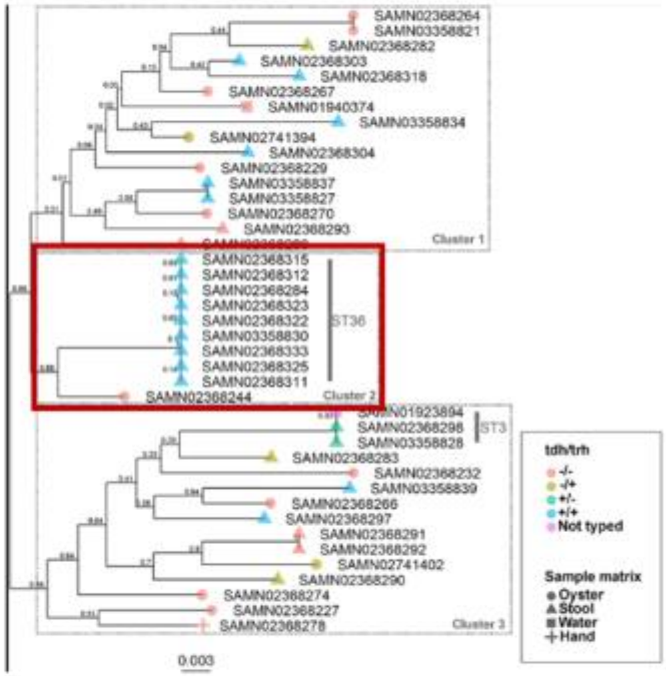
What do you see now?

COMPARING TREE TO ORIGINAL ARTICLE

Phandango



Jesser et al. 2019



What is missing?

SEQUENCE METADATA

- Includes information about each sample
- Provides context

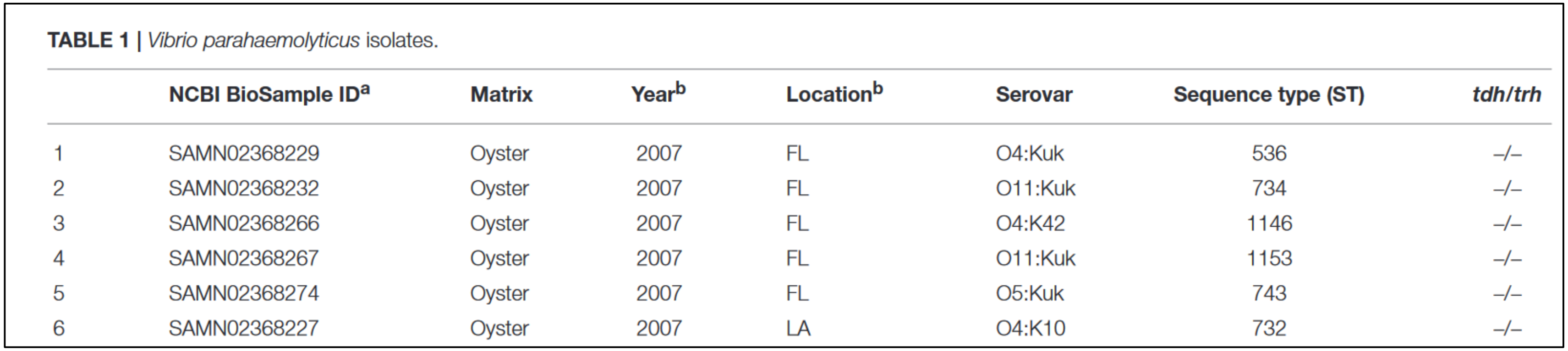
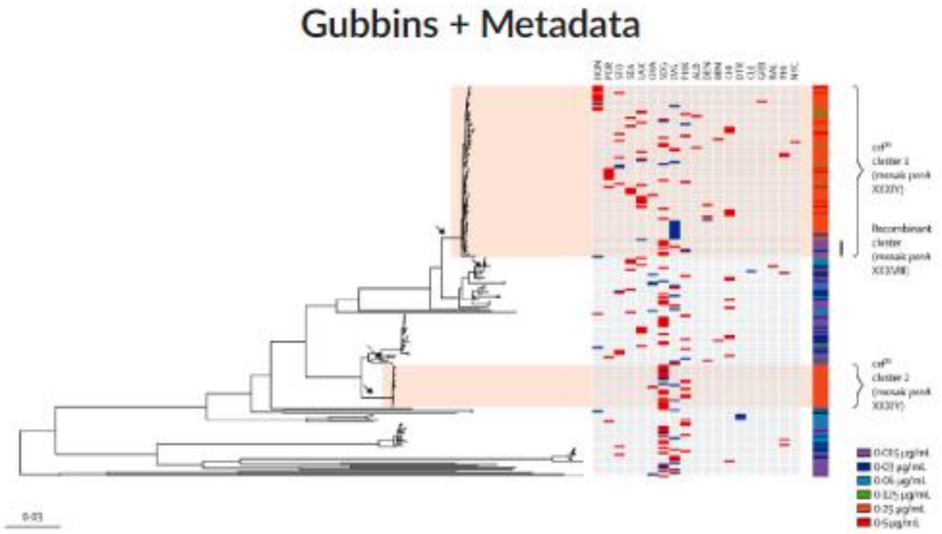


TABLE 1 | *Vibrio parahaemolyticus* isolates.

	NCBI BioSample ID[a]	Matrix	Year[b]	Location[b]	Serovar	Sequence type (ST)	*tdh/trh*
1	SAMN02368229	Oyster	2007	FL	O4:Kuk	536	–/–
2	SAMN02368232	Oyster	2007	FL	O11:Kuk	734	–/–
3	SAMN02368266	Oyster	2007	FL	O4:K42	1146	–/–
4	SAMN02368267	Oyster	2007	FL	O11:Kuk	1153	–/–
5	SAMN02368274	Oyster	2007	FL	O5:Kuk	743	–/–
6	SAMN02368227	Oyster	2007	LA	O4:K10	732	–/–

ADDING METADATA

- Specific file format requirements
- Limited visualizations
- Less customization
- Annotations may cost extra



For more information on adding metadata and input formats visit:
https://github.com/jameshadfield/phandango/wiki/Input-data-formats#metadata

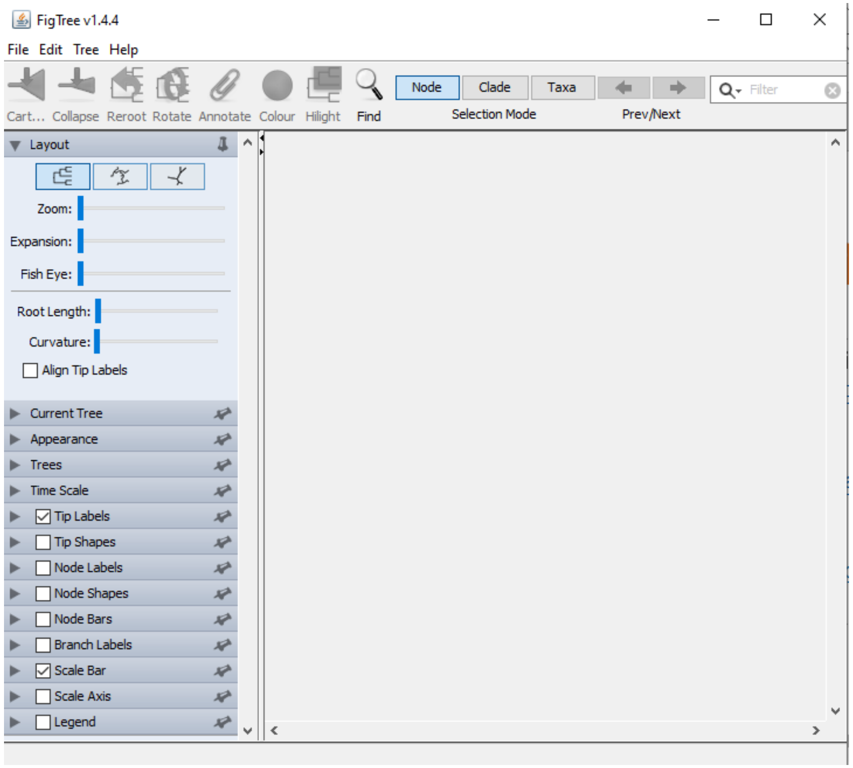
OTHER TREE VIEWING TOOLS

1. ITOL: Interactive Tree of Life
https://itol.embl.de/

2. Auspice
https://auspice.us/

3. FigTree
https://github.com/rambaut/figtree

4. ggTree
https://github.com/YuLab-SMU/ggtree



□ Align Tip Labels
☑ Tip Labels
□ Tip Shapes
□ Node Labels
□ Node Shapes
□ Node Bars
□ Branch Labels
☑ Scale Bar
□ Scale Axis
□ Legend

HELPFUL RESOURCES

Building a phylogenetic tree
https://www.khanacademy.org/science/ap-biology/natural-selection/phylogeny/a/building-an-evolutionary-tree

Phylogenetic algorithms and applications
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7123334/

Overview of phylogenetic tree construction
https://yulab-smu.top/treedata-book/chapter1.html

NEXT WEEK…

Introduction to tree visualization and annotation using ggtree (In Rstudio)
Objectives:

To install ggtree, start R and enter:

```
BiocManager::install(“ggtree”)
```

More information found on the ggtree github, and:
https://bioconductor.org/packages/release/bioc/html/ggtree.html

THANK YOU FOR ATTENDING!

Please make sure to fill out the Exit Survey at https://docs.google.com/forms/d/e/1FAIpQLSeAIQ9WQGEbApPK-EMsicyRgO3TEo-hBUmrDjLqbXWnCjuKOQ/viewform

We value your feedback!

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