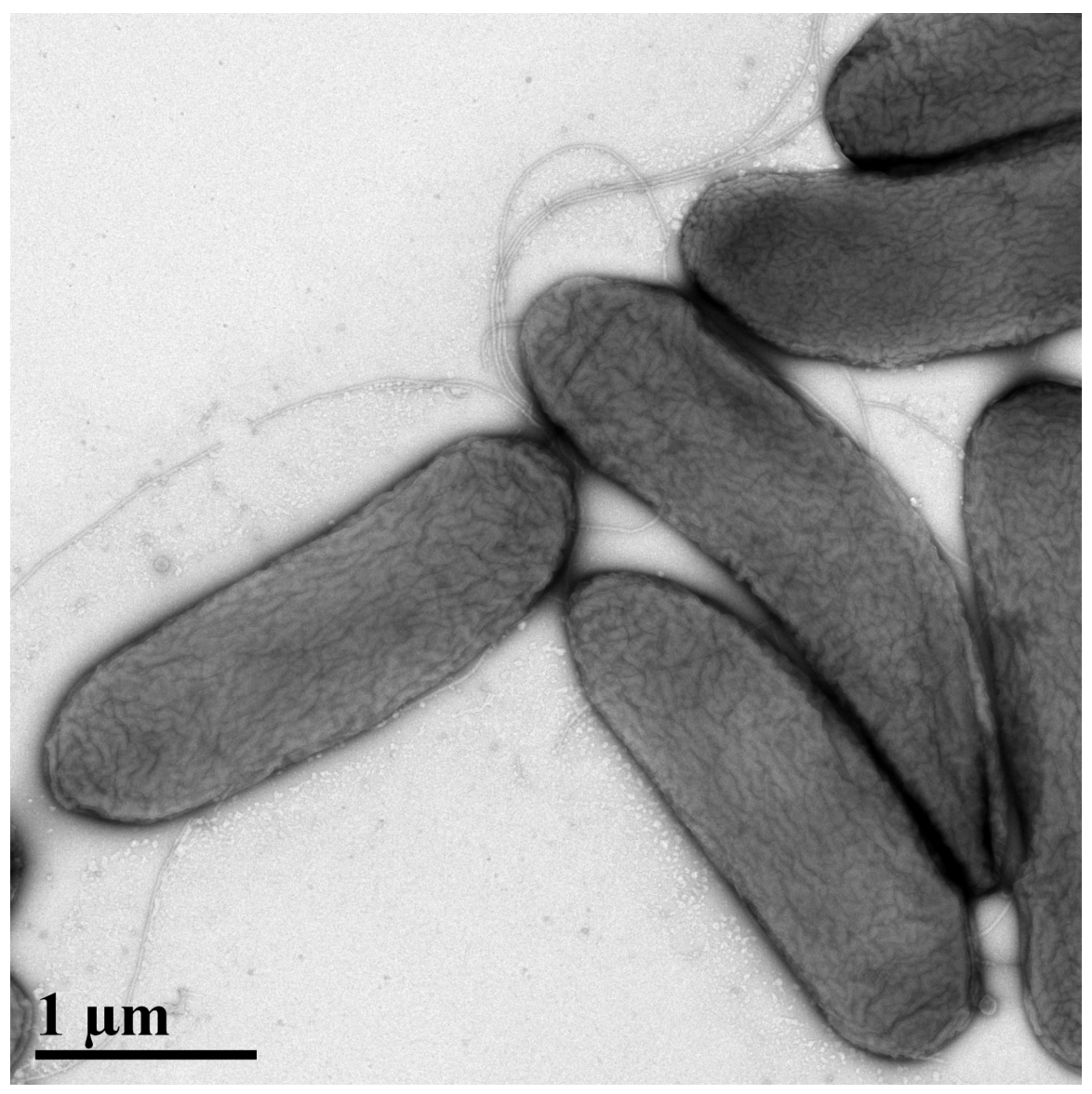
Bacterial Identification II (Afternoon session)

Sujan Timilsina and Jeffrey B. Jones

National Plant Diagnostic Network - 2019

Department of Plant Pathology

University of Florida

National Plant Diagnostic Network - 2019 1

NPDN - 2019

Bacterial Characterization 3

Objective 3

Datasets 3

Software required 3

Sequence Alignment 3

Phylogenetic Analysis 3

16S sequence analysis 4

MLSA genes and datasets 5

Analysis 6

Whole genome sequence comparison 6

Average Nucleotide Identity (ANI) 7

In-Silico DNA-DNA hybridization 7

Comparative Genomics 7

*Xanthomonas* case study 8

16S sequence analysis 8

Hrp genes in *Xanthomonas* 8

MLSA 8

ANI and Insilico DDH 9

# Bacterial Characterization

## Objective

The objective of this workshop is to provide hands-on experience with the sequence based approach used in characterizing and identifying different bacterial strains.

## Datasets

The datasets used in this workshop are available in the following link: <https://github.com/sujan8765/NPDN>

## Software required

Please download the following software:

MEGA - (<https://www.megasoftware.net/>)

## Sequence Alignment

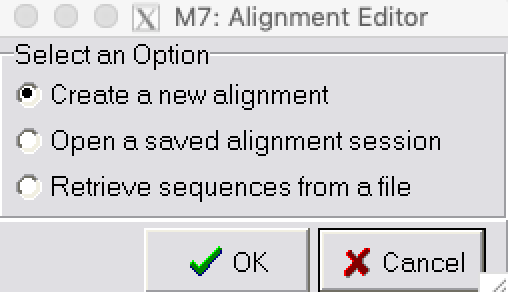
Sequence alignment is the method of arranging nucleotide and amino acid sequences to identify similarity between the sequences. Number of different sequence alignment algorithms and software are available. These can be accessed via webpage or are available through different software.

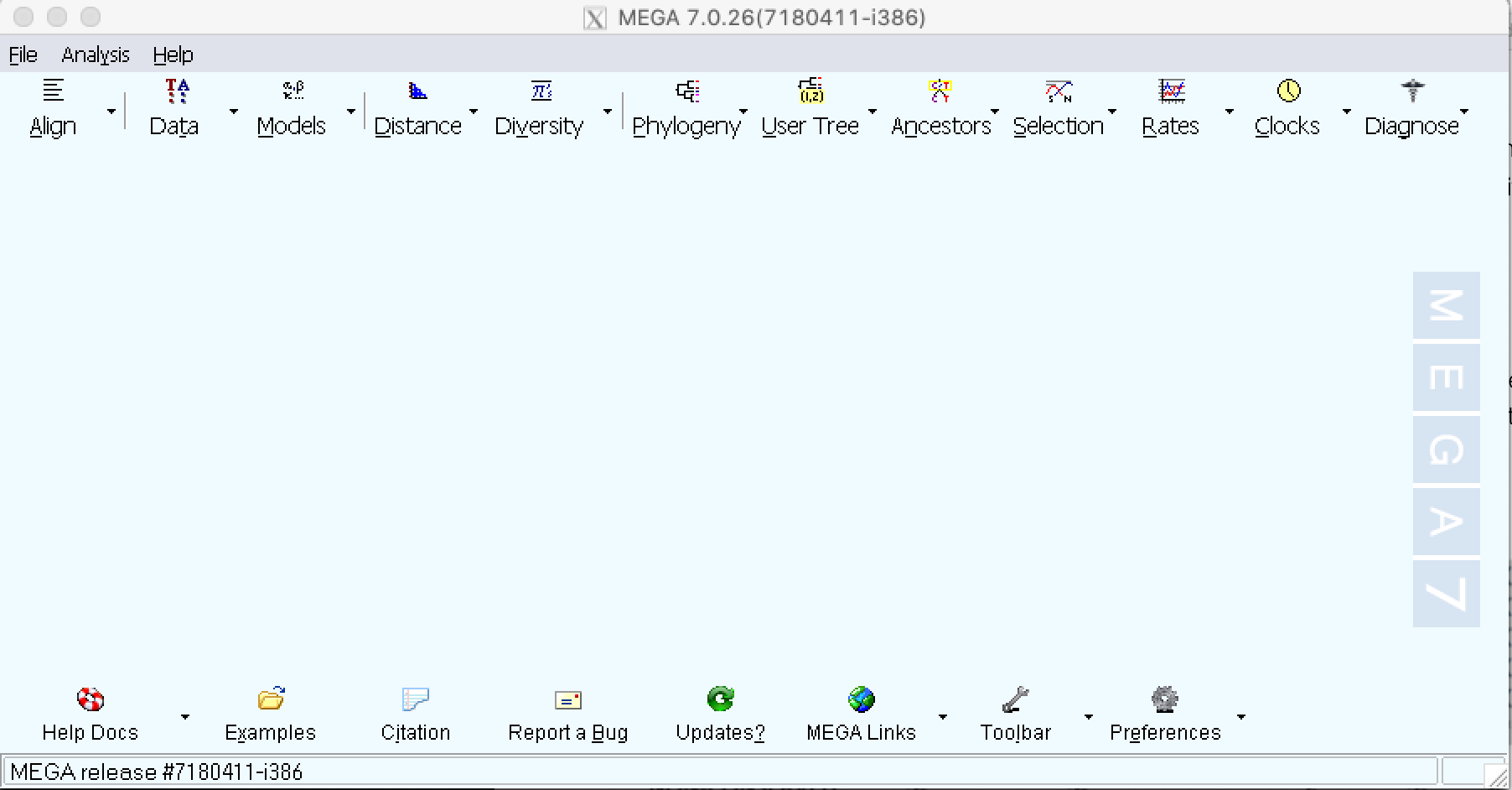
## Phylogenetic Analysis

Evolutionary models are selected using aligned sequences for further phylogenetic analyses. Software like jModelTest and MrModelTest are used to determine the best models. Commonly used methods to construct trees are Maximum Likelihood and Bayesian methods. Maximum likelihood (ML) provides probability of sequences in a model for the tree and tree with highest probability is referred. ML trees are computationally intense and provides ability to choose evolutionary models unlike in distance-based trees. Similarly, Bayesian methods are based on maximum likelihood with posterior probabilities and uses complex models such as Markov Chain Monte Carlo methods. We will be constructing maximum likelihood phylogenetic trees.

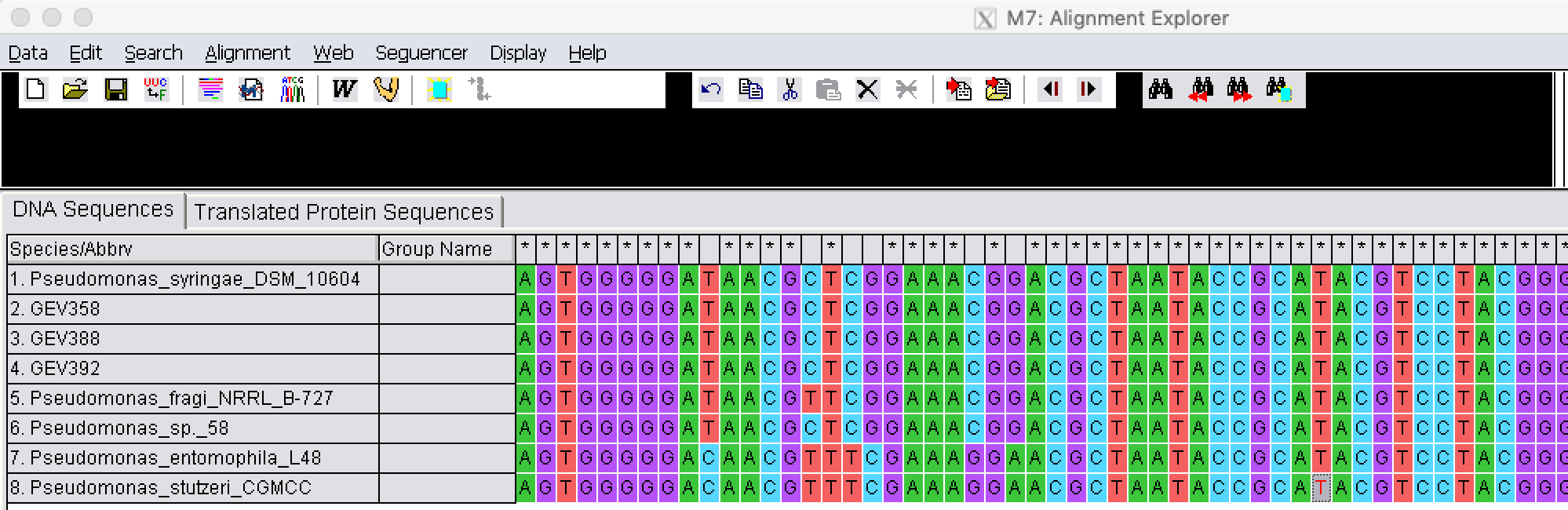
## 16S sequence analysis

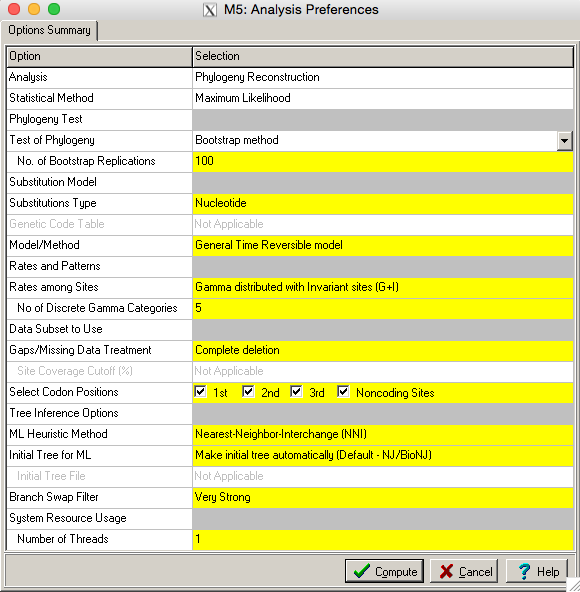
16S sequences are ribosomal genes that are consistently used for preliminary characterization primarily because i) it is present in all bacteria, ii) function of the gene remain unchanged thus providing a more accurate measure of microbial evolution and iii) the size of genomic fragment of around 1500 nucleotides provide a significant information for characterization and initial identification purposes. Universal and broad range primers can be used for amplifying and sequencing the gene for comparisons with reference sequences.

Reference sequences could be downloaded from NCBI (<https://www.ncbi.nlm.nih.gov/>). Information on prokaryotic type strains and publicly available sequences could be accessed from (<https://www.ezbiocloud.net>).



1. A sample 16S sequences from representative strains are available from [this link](https://github.com/sujan8765/NPDN/tree/master/Pseudomonas_sample_data).
2. Open the MEGA software on your computer.
3. Click ‘Align’ on top left and ‘Edit/Build Alignment’.
4. You can either ‘Create a new alignment’ and copy the sequences or ‘Retrieve sequences from a file’ and upload the sequences.
5. Click Alignment on the title menu or the highlighted options for sequence alignment methods ‘ClustalW’ and ‘Muscle’.



1. Using default settings click ‘Compute’ and save the alignment in mega format from the dropdown menu: ’Data > Export alignment > MEGA Format’ and save it as ‘16S.meg’.
2. Click ‘Phylogeny > Construct/Test Maximum Likelihood Tree’. Open the ‘16S.meg’ file.
3. Bootstrap test of phylogeny could be added from the dropdown menu, and different models/methods with variable rates can be used for constructing phylogenetic tree. The models may vary depending on the sequences and the likelihood of model that could fit the phylogeny could be determined using jModelTest software (<https://github.com/ddarriba/jmodeltest2>). For the session, please use General Time Reversible model with Gamma distributed Invariant sites.
4. Save the image in different formats from the dropdown ‘Image’ menu in the title bar.

### Multilocus Sequence Analysis/Typing (MLSA/T)

Multilocus sequence analysis/typing (MLSA/T) was proposed in late 1990s that uses partial gene sequences of several housekeeping genes. First proposed as sequence typing to characterize human pathogen *Neisseria* *meningitidis,* MLSA approach uses nucleotide sequences in multiple loci of bacterial genome that encode housekeeping genes and are analyzed based on the sequences. MLST approach converts each sequence to numerical allele type that can be employed for population studies. With increased sites spread around the genome, MLSA/T provides improved analytical methods for epidemiological and population studies.

## MLSA genes and datasets

MLSA are relatively genus specific. Different multilocus sequence schemes have proposed multiple sets of housekeeping genes that could be amplified for sequence comparisons. Some independent databases with MLSA genes of different plant pathogenic bacteria are available online. However, most of the sequences are publicly available through NCBI (<https://www.ncbi.nlm.nih.gov/>). PAMDB, Plant Associated and Environmental Microbes Database (<http://genome.ppws.vt.edu/cgi-bin/MLST/home.pl>) hosts MLSA datasets of common plant pathogenic bacteria: *Xanthomonas, Pseudomonas, Ralstonia, Acidovorax* and *Clavibacter michiganensis*. Similarly, MLST datasets of *Xylella fastidiosa* and *Candidatus* Liberibacter solanacearum can be accessed through pubMLST (<https://pubmlst.org/databases/>).

## Analysis

1. Create an user account in [www.pamdb.org](http://www.pamdb.org). Using your user credentials and during login select ‘Pseudomonas’ as organism. Download the four housekeeping genes *gap1, gltA, gyrB,* and *rpoD* sequenced from representative GEV388 strain from the link. Click BLAST and copy sequences individually and click search. Provide the list of name of strain with highest sequence similarity.
   1. *gap1* - \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
2. *gyrB* - \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
3. *gltA* - \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
4. *rpoD* - \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
5. There are four housekeeping genes from a *Pseudomonas* *floridensis* strain are included separately in fasta format [here](https://github.com/sujan8765/NPDN/tree/master/Pseudomonas_sample_data). Please compare these genes separately to the NCBI database using blast (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) and identify the sequences (other than *P. floridensis*) with the highest similarity. Please fill the accession number and species name of the sequence listed after *P. floridensis*:
   1. *gap1* - \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
   2. *gyrB* - \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
   3. *gltA* - \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
   4. *rpoD* - \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
6. Now open the ‘Pseudomonas\_concatenated.fas’ sequence file ([link](https://github.com/sujan8765/NPDN/tree/master/Pseudomonas_sample_data)) that includes the concatenated sequence of the *Pseudomonas floridensis* strain along with other reference sequences. For each gene, the single gene sequences were individually aligned (same steps as for 16S sequence alignment) and concatenated in the same order for all sequences. Please align the concatenated sequences one more time, save the file in mega format and construct a maximum likelihood phylogenetic tree using the same method as 16S sequence analysis.
7. ‘Loropetali\_sample\_data.meg’ is an aligned mega formatted file with concatenated sequences for four housekeeping genes of *Pseudomonas amygdali* pv*. loropetali* and reference strains. Construct a maximum likelihood tree using the methods same as above.

## Whole genome sequence comparison

Ability to sequence bacterial genome rapidly at low-cost has revolutionized bacterial characterization and taxonomy. The first plant pathogenic bacteria sequenced was *Xylella fastidiosa.* Complete whole genomes or draft genomes of more than 120 different plant pathogenic bacterial species are publicly available in NCBI database. Different sequencing platforms are used to obtain the nucleotide sequences of bacteria. These raw sequence outputs are assembled using specific genome assembly pipelines like Spades, Velvet, or via commercial software like CLC Genomics Workbench or Geneious. The whole genomes from the bacterial strain can be used to compare with reference strains to characterize and confirm the taxonomy. Whole genome based sequence comparisons like average nucleotide identity and *insilico* DNA-DNA hybridization are the most common methods for pairwise comparisons of the genomes.

## Average Nucleotide Identity (ANI)

Average nucleotide identity (ANI) is the measure of nucleotide similarity between the two genomes. ANI measures the sequence identity between the genomic pairs that can be based on the whole genomes or between the coding regions of the two genomes. ANI > 95% suggests that the two genomes likely belong to the same species and could be further verified based on comparative genomics. ANI could be calculated between the genomes in the IMG/JGI platform, ANI calculator (<http://enve-omics.ce.gatech.edu/ani/>) or standalone software like jSpecies (<https://imedea.uib-csic.es/jspecies/download.html>).

1. Assembled genome of *Pseudomonas floridensis* strain GEV388T can be downloaded from [the link](https://github.com/sujan8765/NPDN/tree/master/Pseudomonas_sample_data).
2. Genomes from representative Pseudomonas species that were closely related with GEV388 based on the MLSA are also provided in the same folder.
3. Upload the genomes to ANI calculator (<http://enve-omics.ce.gatech.edu/ani/>) and calculate pairwise ANI among these species/strains.

## In-Silico DNA-DNA hybridization

Genome-to-Genome Distance Calculator (GGDC) can be used for calculating DNA-DNA hybridization based on sequence alignments.

1. Upload the assembled genome of Pseudomonas floridensis strain GEV388T to <http://ggdc.dsmz.de/ggdc.php> as query genome.
2. Upload the references genomes downloaded from ([reference Pseudomonas genomes](https://github.com/sujan8765/NPDN/tree/master/Pseudomonas_sample_data)) to <http://ggdc.dsmz.de/ggdc.php> as reference genomes.
3. Under default setting as BLAST+ as the recommended alignment submit for analysis.
4. Check the output in your email that gives you an output about genetic distance and DNA-DNA hybridization estimate.

## Comparative Genomics

Along with assembled genomes, annotated genomes and gene clusters are compared to provide additional information on bacterial genome similarity and differences. Different annotation pipelines are used to annotate prokaryotic genomes. NCBI uses Prokaryotic Genome Annotation Pipeline to annotate genomes. Similarly, IMG/JGI (Integrated Microbial Genomics/Joint Genome Institute <https://img.jgi.doe.gov/>) uses different databases like Clusters of Orthologous Genes (COGs), Kyoto Encyclopedia of Genes and Genomes (KEGG) and other additional databases for genome annotation. Rapid Annotation using Subsystem Technology (RAST), and Rapid prokaryotic genome annotation (PROKKA) are some of the web-based and standalone tools for bacterial genome annotation. For plant pathogenic bacteria, pathogenicity associated genes are used for comparative purposes. Type III secretion system, effector profiles, and cell-wall degrading enzymes are the common genes and genomic fragments used to compare between strains and species.

# *Xanthomonas* case study

In the above description, we focused on approach when a new species was characterized based on sequence information. We will now focus on a case study where two different species of *Xanthomonas* were merged ([link to the manuscript](https://ijs.microbiologyresearch.org/content/journal/ijsem/10.1099/ijsem.0.003104)). Recently, *Xanthomonas cynarae* and *X. gardneri* were merged as a same species and *X. cynarae* was used as the species name as it was described prior to *X. gardneri*. The strains of *X. cynarae* were reported pathogenic on artichoke and *X. gardneri* causes bacterial spot disease in tomato and pepper. Pathogenicity assay on tomato, pepper, and artichoke identified that the strains of *X. cynarae* and *X. gardneri* interact distinctly with the hosts. *X. cynarae* strains didn’t cause disease in pepper and *X. gardneri* was not pathogenic on artichoke. We used sequence based analysis and comparative genomics to merge the species. However, due to distinct hosts, the previous species names were retained as pathovars of *X. cynarae* as *X. cynarae* pv. *cynarae* and *X. cynarae* pv*. gardneri*.

16S sequence analysis

1. Please download the aligned 16S sequences of *X. cynarae* along with other reference *Xanthomonas* strains (link to sequence file). The sequences were compiled from ezbiocloud and NCBI sequence database.
2. Using methods previously described, upload the sequences to MEGA, verify alignment, and construct a Maximum likelihood phylogenetic tree. Please run bootstrap phylogeny test with at least 100 repeats.

Hypersensitive response and pathogenicity (Hrp) genes in *Xanthomonas*

Type III Secretion System associated HrcN genes are used for strain characterization and diagnosis in *Xanthomonas*. Protocols for diagnosis using real-time and isothermal PCR techniques are already available. Please download the hrcN sequence alignment from [here](https://github.com/sujan8765/NPDN/tree/master/Xanthomonas_sample_data) and run a phylogenetic analysis on how the *Xanthomonas* species correlate with each other. The alignment was used to construct phylogenetic tree for strains sampled in Ethiopia (publication link: [Kebede et al. 2014](https://link.springer.com/article/10.1007/s10658-014-0497-3))

MLSA

1. Six housekeeping genes suggested by [Almeida et al. 2010](https://apsjournals.apsnet.org/doi/abs/10.1094/PHYTO-100-3-0208) (*fusA, gapA, gltA, gyrB, lacF, and lepA)* were sequenced from representative strains using respective primers. The aligned and concatenated sequences can be downloaded from [this link](https://github.com/sujan8765/NPDN/tree/master/Xanthomonas_sample_data). Similar to method described as above to construct a maximum likelihood tree using MEGA.
2. Would you consider *X. hortorum* as the same species to *X. cynarae* and *X. gardneri*?

ANI and Insilico DDH

* 1. Upload the genomes to the web based platform (<http://enve-omics.ce.gatech.edu/ani/>) to calculate ANI or to jSpecies to calculate ANI based on blast and mummer.
  2. Similar to the methods described above, please run insilico DNA-DNA hybridization using genome-to-genome distance calculator (<http://ggdc.dsmz.de/ggdc.php>).

Discussion points

1. Would you consider *X. gardneri, X. cynarae,* and *X. hortorum* as same species or different? Why or why not?
2. What other analysis you could do to conclude your study?

Datasets used in this workshop are associated with following publications:

Timilsina, S., Kara, S., Jacques, M.A., Potnis, N., Minsavage, G.V., Vallad, G.E., Jones, J.B. and Fischer-Le, M.S., 2019. Reclassification of *Xanthomonas gardneri* (ex Šutič 1957) Jones et al. 2006 as a later heterotypic synonym of *Xanthomonas cynarae* Trébaol et al. 2000 and description of *X. cynarae* pv*. cynarae* and *X. cynarae* pv*. gardneri* based on whole genome analyses. *International journal of systematic and evolutionary microbiology*, *69*(2), pp.343-349.

Kebede, M., Timilsina, S., Ayalew, A., Admassu, B., Potnis, N., Minsavage, G.V., Goss, E.M., Hong, J.C., Strayer, A., Paret, M. and Jones, J.B., 2014. Molecular characterization of Xanthomonas strains responsible for bacterial spot of tomato in Ethiopia. *European journal of plant pathology*, *140*(4), pp.677-688.

Harmon, C.L., Timilsina, S., Bonkowski, J., Jones, D.D., Xiaoan, S., Vallad, G.E., Sepulveda, L.R., Bull, C., and Jones, J.B., 2018. Bacterial Gall of *Loropetalum chinense* caused by *Pseudomonas amygdali* pv*. loropetali* pv. nov. *Plant Disease*. Pp.799-806.

Timilsina, S., Minsavage, G.V., Preston, J., Newberry, E.A., Paret, M.L., Goss, E.M., Jones, J.B., and Vallad, G.E. 2017. *Pseudomonas floridensis* sp*.* nov., a bacterial pathogen isolated from tomato. *International journal of systematic and evolutionary microbiology*. Pp. 64-70.

For any feedbacks and comments, please feel free to connect

email: [sujan.timilsina@ufl.edu](mailto:sujan.timilsina@ufl.edu)

twitter: [@sujantimilsina](https://twitter.com/sujantimilsina)

web: <https://www.sites.google.com/site/sujantimilsina/>