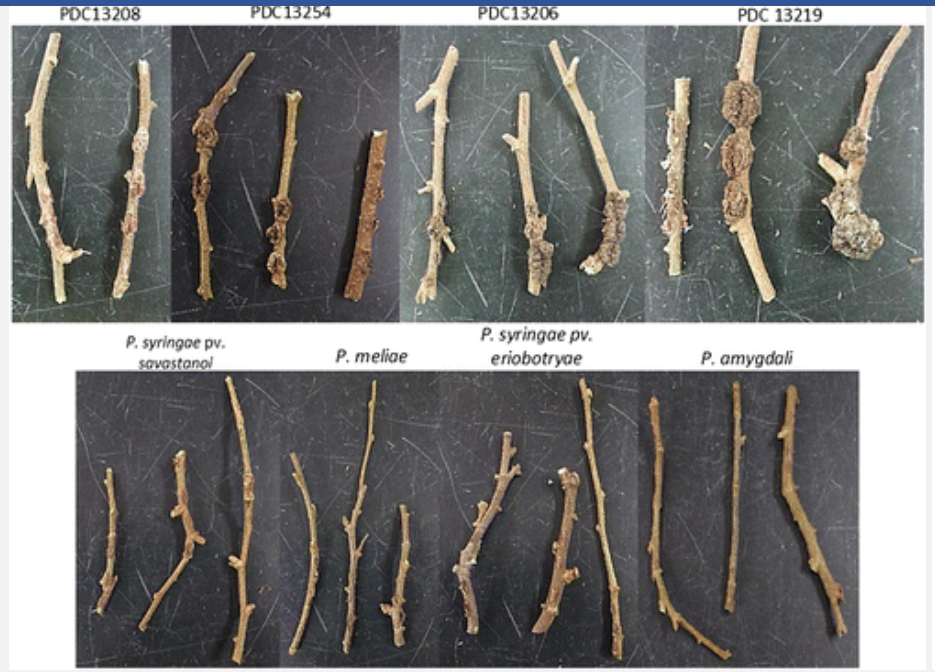
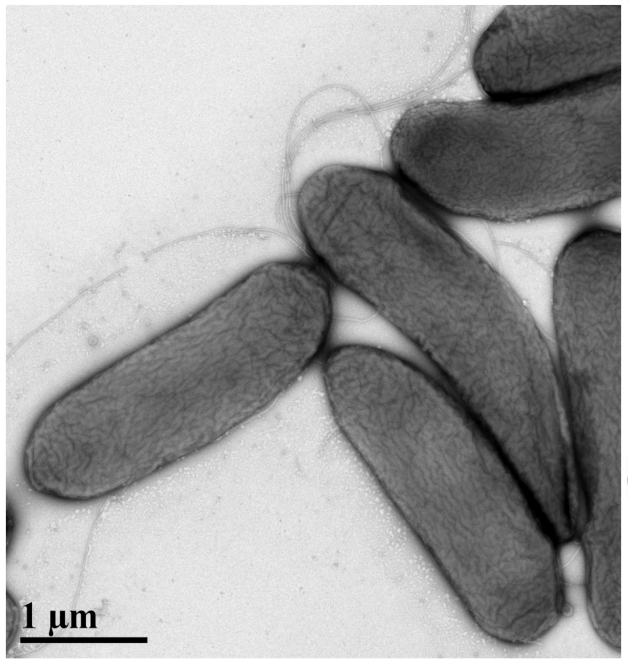
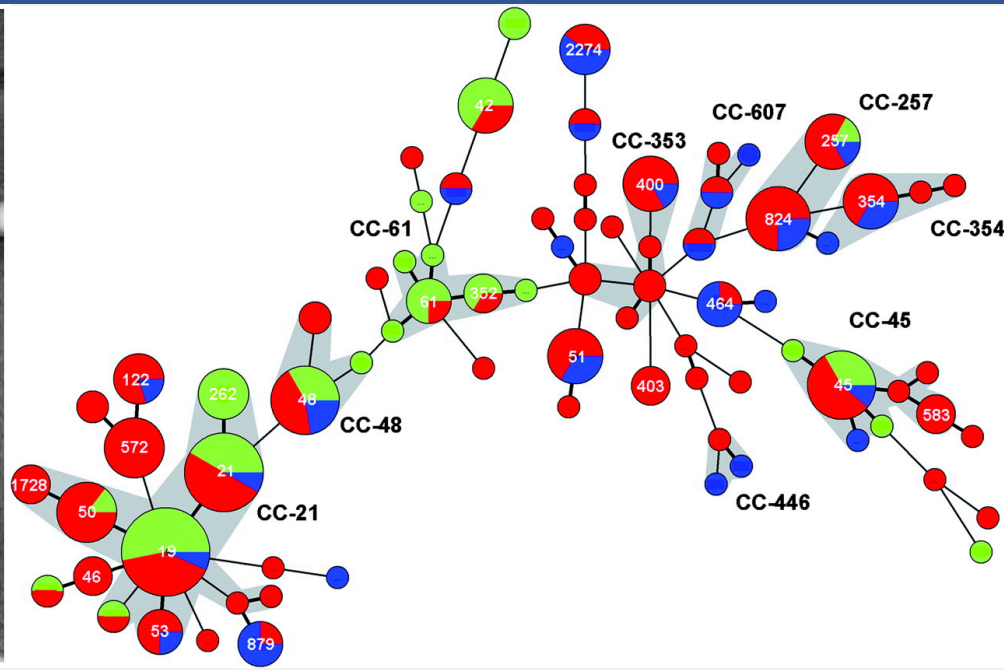






Bacterial Identification II (Afternoon Session)

Sujan Timilsina, Ph.D.
Department of Plant Pathology
University of Florida

Workshop Outline

- Sequence based prokaryotic diagnostics
  - 16S sequence analysis
  - Multilocus sequence analysis
  - Whole genome sequence analysis
- Submitting sequences to databases: NCBI
- Translational tools

Bacterial diagnostics

- In morning session
  - Isolating bacteria,
  - Phenotypic methods of disease diagnostics,
  - Using fatty acid composition for pathogen characterization,
  - Carbon source utilization and using strain properties to discern metabolic fingerprinting using Biolog.
  - DNA-DNA hybridization

Sequence based bacterial identification

- Preliminary bacterial identification involves phenotypic characterization.
- We will be focusing on sequence based bacterial identification, characterization and taxonomy.
- 16S ribosomal gene sequences are commonly used for preliminary characterization.

16S sequences

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 Primers for 16S

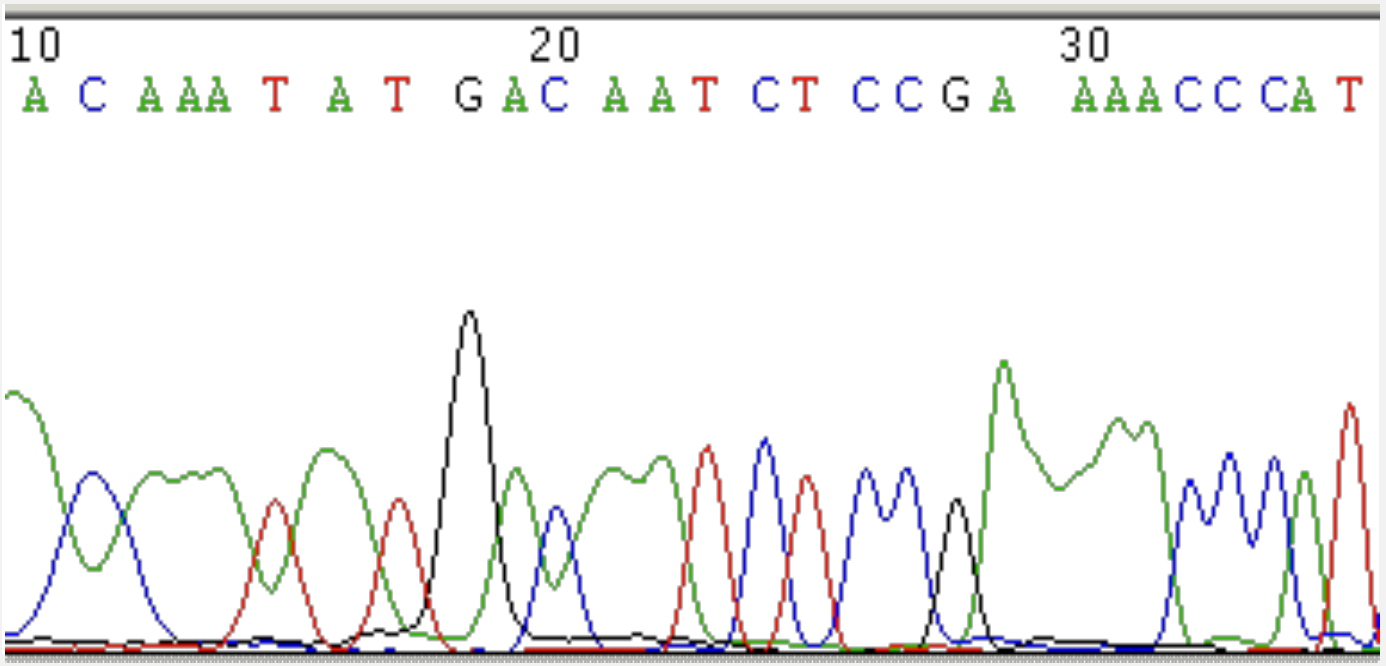
- Use of 16S was proposed in 1977.
- Weisberg et al. 1991 used 16S ribosomal DNA for phylogenetic study.

Primer name	Sequence (5'-3')
8F	AGA GTT TGA TCC TGG CTC AG
U1492R	GGT TAC CTT GTT ACG ACT T
928F	TAA AAC TYA AAK GAA TTG ACG GG
336R	ACT GCT GCS YCC CGT AGG AGT CT
1100F	YAA CGA GCG CAA CCC
1100R	GGG TTG CGC TCG TTG
337F	GAC TCC TAC GGG AGG CWG CAG
907R	CCG TCA ATT CCT TTR AGT TT
785F	GGA TTA GAT ACC CTG GTA
805R	GAC TAC CAG GGT ATC TAA TC
533F	GTG CCA GCM GCC GCG GTA A
518R	GTA TTA CCG CGG CTG CTG G
27F	AGA GTT TGA TCM TGG CTC AG
1492R	CGG TTA CCT TGT TAC GAC TT

Consensus sequence

- Consensus sequences from forward and reverse sequence.

- Chromatogram images provide information on nucleotide peaks to verify the sequence.



Database available

- NCBI: National Center for Biotechnology Information

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

- Ezbiocloud:

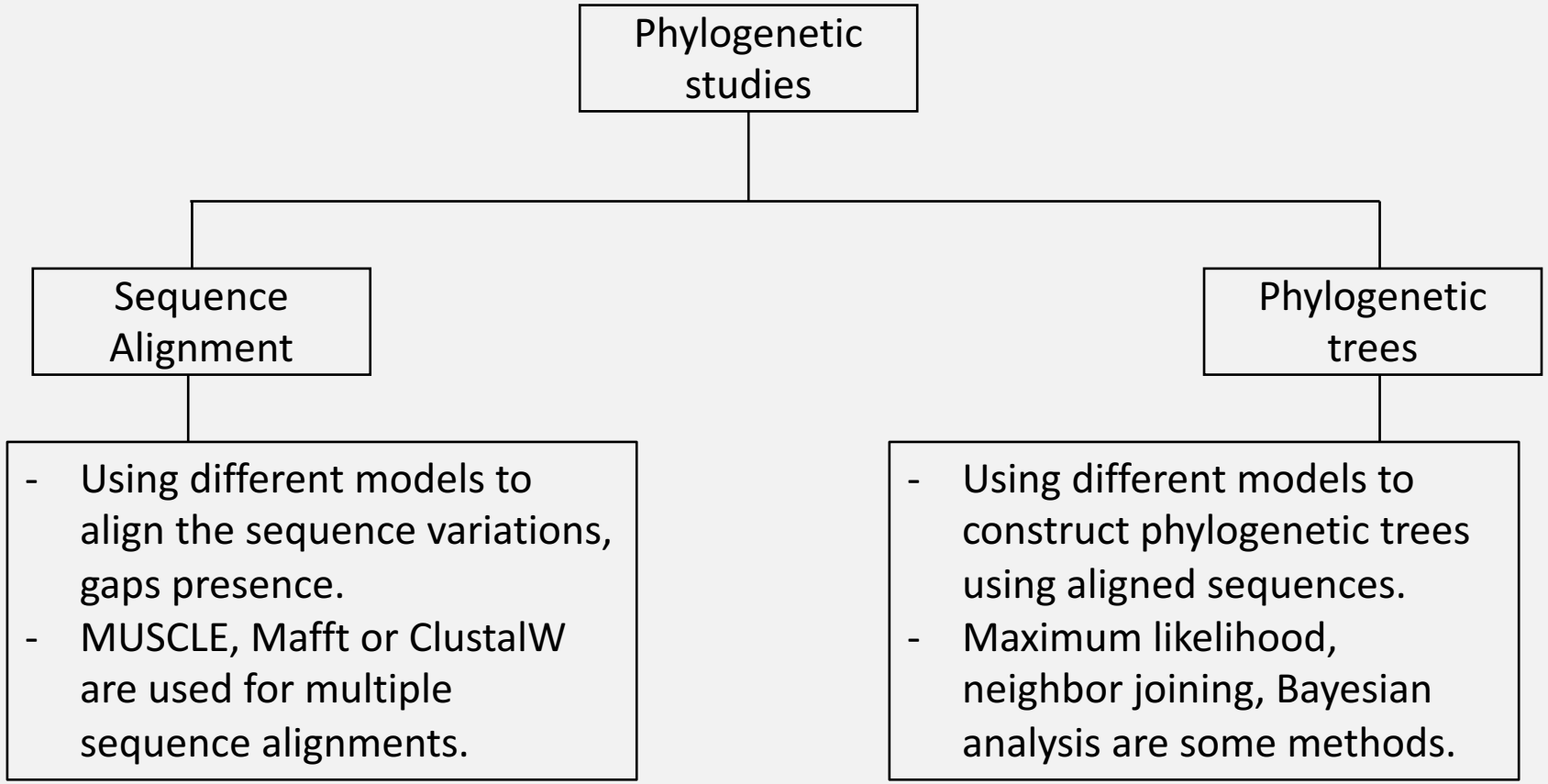
  https://www.ezbiocloud.net/

Phylogenetic analysis

- Sample data from *Pseudomonas amygdali* pv. *loropetali* and *P. floridensis*

- Link to the repository: https://github.com/sujan8765/NPDN

Phylogenetic studies

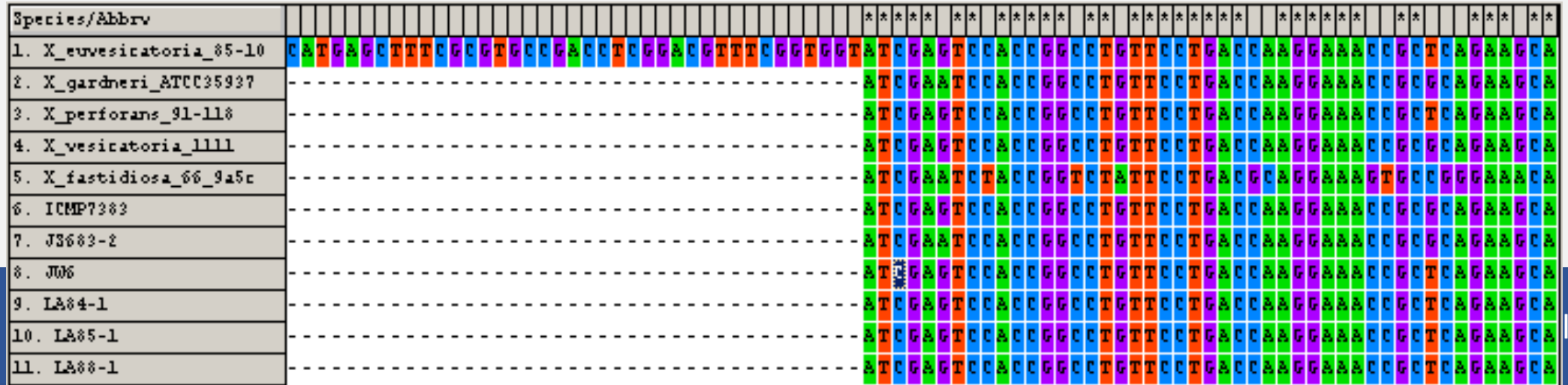


Sequence Alignments

It is a way of arranging sequence information from DNA, RNA or protein to identify regions of similarity or differences, that may be a consequence of functional, structural or evolutionary relationships between the sequence.

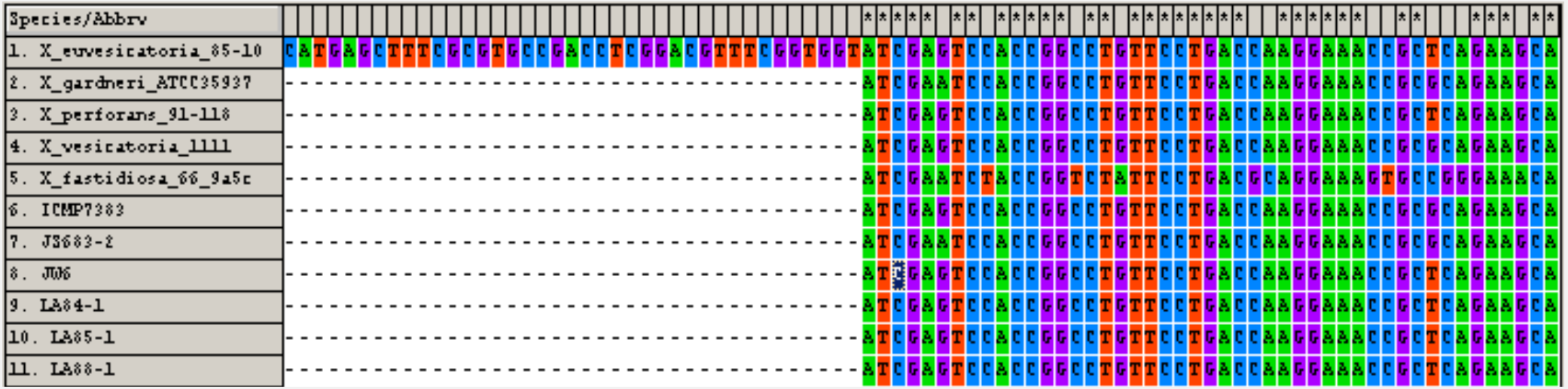
Wrong alignment results in wrong interpretation from data.

Different alignment models available like MUSCLE, Mafft and ClustalW.

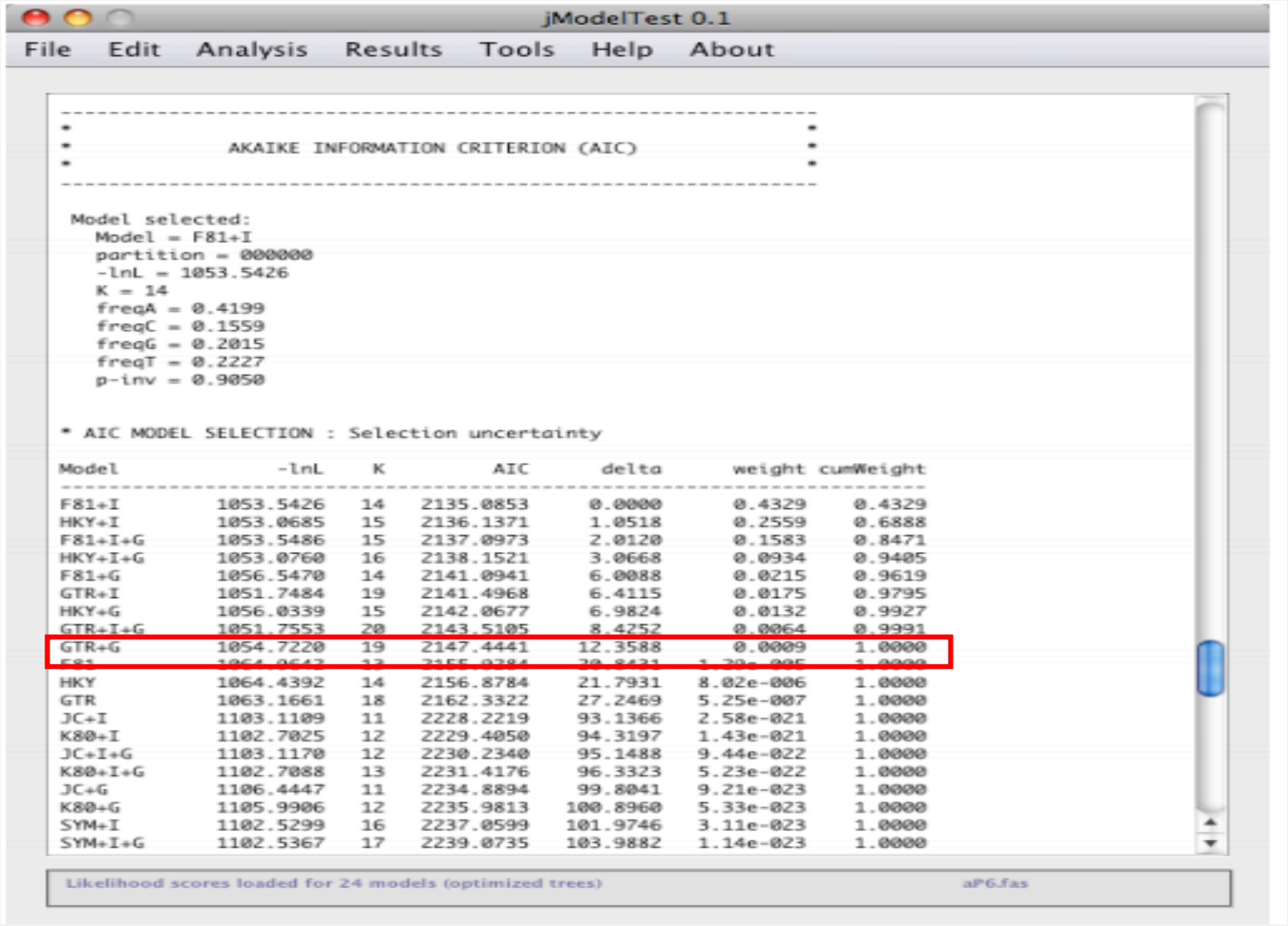


Creating Sequence Alignment

- Copy your sequences into MEGA
- Click the ClustalW or Muscle.
- Depending upon your input variable sequences will be aligned accordingly.
- Remove the extra ‘flanking’ sections.



Model Selection



Maximum likelihood

- Statistical method for estimating unknown parameters of probability model.

Neighbor Joining

- Bottom-up clustering method for creating phylogenetic trees.

Maximum parsimony

- Predicts evolution by minimizing the number of steps required to generate the observed variation.

Bayesian Inference

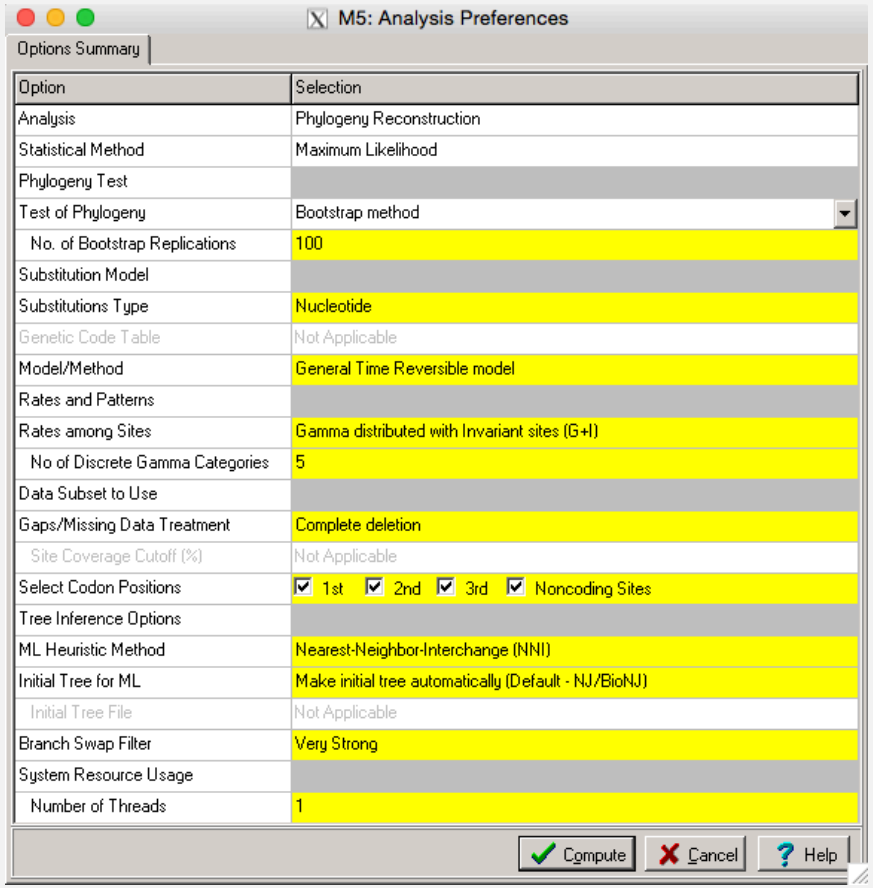
- Uses likelihood function to create a posterior probability of trees.

Phylogenetic inference

- Select model and test for Phylogeny

Bootstrap test of Phylogeny

Procedure of resampling the sites and the subsequent tree reconstruction is repeated several hundred times, and the percentage of times each interior branch is given a value is noted and the values are known as bootstrap values. Higher bootstrap value means higher confidence in the tree topology.



Results

- Compare outputs from different models
  - with or w/o bootstrap test.
  - Anterior and posterior probabilities
  - Different models of nucleotide substitution.

- Create a consensus tree.

Datasets and Software

- Datasets:

  https://github.com/sujan8765/NPDN

- Software:

  MEGA: https://www.megasoftware.net/

- To open .fna / .fasta – windows (notepad), mac (textedit or Bbedit)

Lets start with 16S

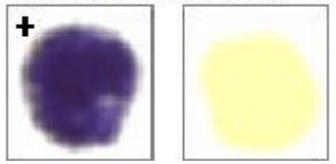
- A bacteria strain GEV388 was isolated in 2011, LOPAT assay suggested similar to *Pseudomonas viridiflava.* (Timilsina et al. 2017. *Pseudomonas floridensis* sp. nov., a bacterial pathogen isolated from tomato)
- Phenotypically characterized using LOPAT test:

  L – Levan production O – Oxidase reaction

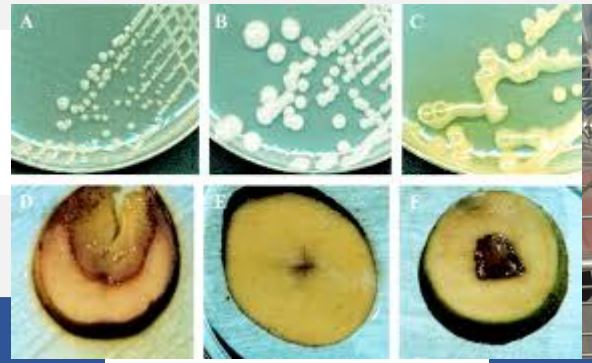
  P – Pectinolytic activity A – Arginine dihydroalse

  T – Hypersensitive reaction on Tobacco

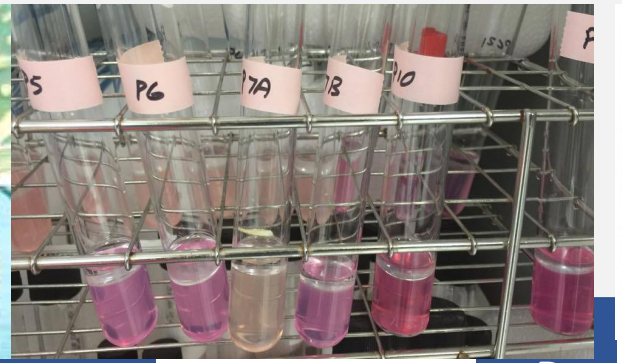
Levan production



Oxidase reaction



Pectinolytic assay



Arginine assay

Tobacco hypersensitivity

16S sequence analyses

- The strain was sequenced using *Pseudomonas* specific primers for 16S sequence.

- The GEV388 strain along with other reference strains can be downloaded from https://github.com/sujan8765/NPDN/tree/master/Pseudomonas_sample_data

- Construct a Maximum Likelihood phylogenetic tree.

Multilocus Sequence Analysis with multiple housekeeping genes

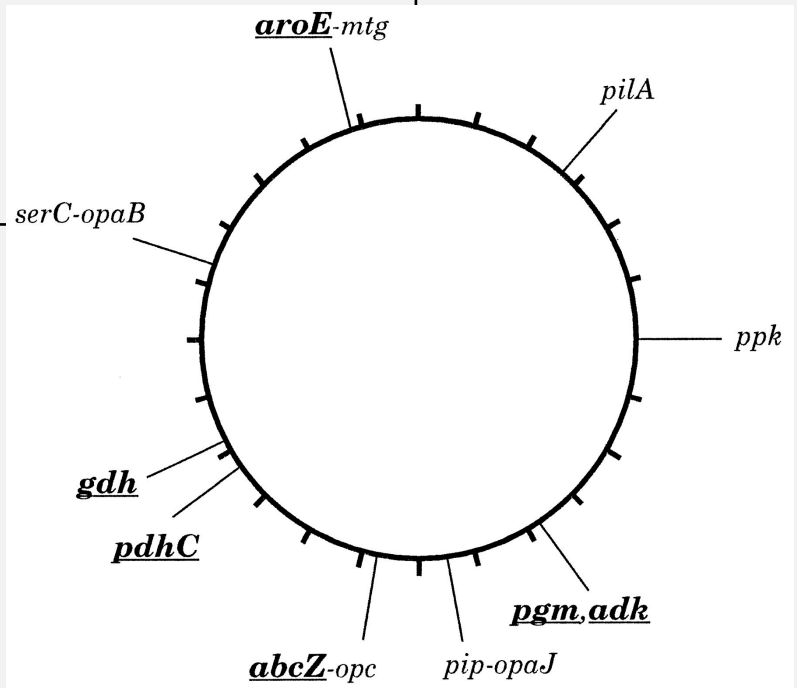
Introduction of MLST/A

- MLST/A was first proposed in 1998 by Maiden and colleagues, for naturally transformable Gram-negative pathogen *Neisserria meningitidis* but has since been applied to many pathogenic species and, more recently, environmental bacteria and eukaryotes.

- The procedure is essentially an updated version of multilocus enzyme electrophoresis (MLEE), which indexes variation within **multiple core metabolic (housekeeping) genes** on the basis of differing electrophoretic mobility of the gene products.

Multilocus Sequence Typing/Analysis (MLST/A)

Multilocus Sequence Typing	Multilocus Sequence Analysis
a method for the genotypic characterization of prokaryotes at the intraspecific level, using the **allelic mismatches** of a small number of housekeeping genes. It is designed as a tool in molecular epidemiology and used for recognizing distinct strains within named species.	a method for the genotypic characterization of a more diverse group of prokaryotes **using the sequences** of multiple protein-coding genes.



MLST/A are DNA sequence based approach for the **unambiguous** characterization of isolates of bacteria and other organisms.

Schemes are based on DNA sequence of typically **4-10 loci** in a bacterial genome that **are under neutral selection**.

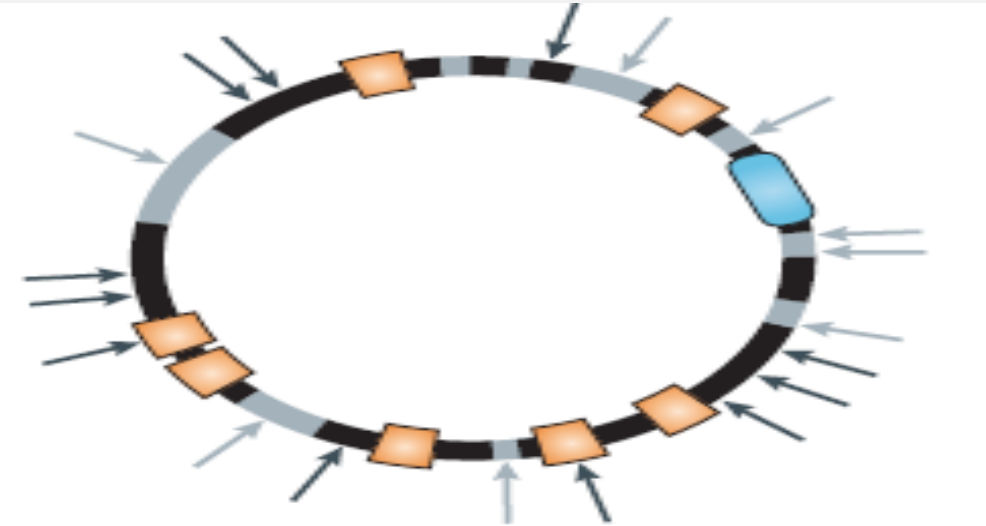
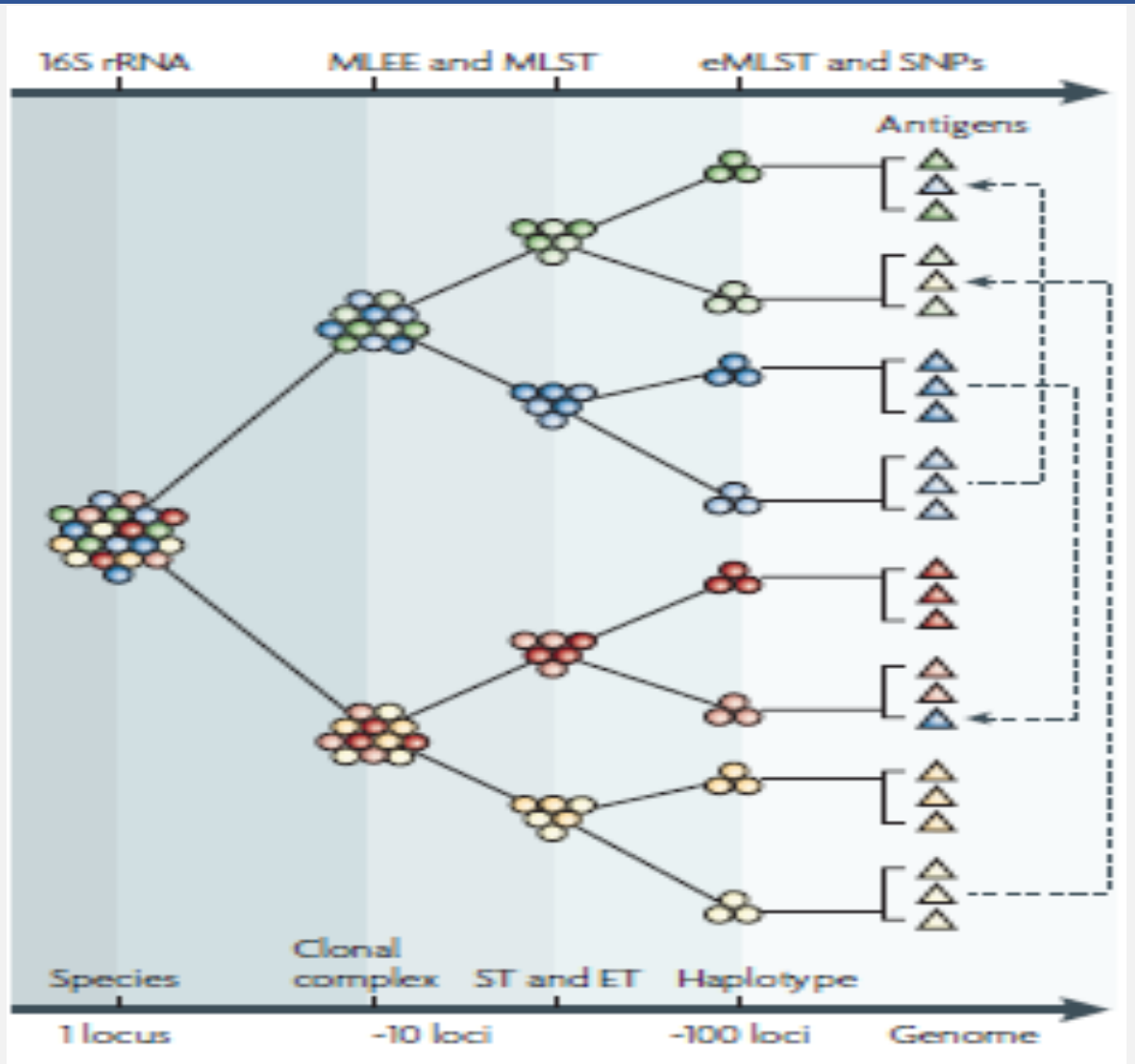


Figure 1 | **Genomic coverage of genetic typing methods.** Shows the core genome, which includes genes that encode proteins which are involved in essential functions, such as replication, transcription and translation, and the dispensable genome, which includes genes that encode proteins that facilitate organismal adaptation. Coverage by 16S ribosomal RNA (rRNA), multilocus sequence typing (MLST) and single-nucleotide polymorphisms (SNPs) is also depicted. For *Neisseria meningitidis*, which has a 2.2 Mb genome[91], the average length of a 16S rRNA gene is ~1.5 Kb[92] and the average length of the MLST loci is ~4 Kb[6]; typing based on 16S rRNA and MLST therefore covers 0.07% and ~0.2% of the *N. meningitidis* genome, respectively. The genome of *Salmonella enterica* serovar Typhi (*S. typhi*) is ~4.8 Mb[93] and SNP gene fragments are present in ~89 Kb[23]; SNP-based typing therefore provides coverage of 2% of the *S. typhi* genome.

MLST is usually applied to strains that belong to well-defined species while MLSA is generally used when species boundaries are not well known and MLSA data can improve species description.

Different levels of resolution within a population as identified by different typing methods



Primers

- For *Xanthomonas*, degenerate primers are designed for selected loci.

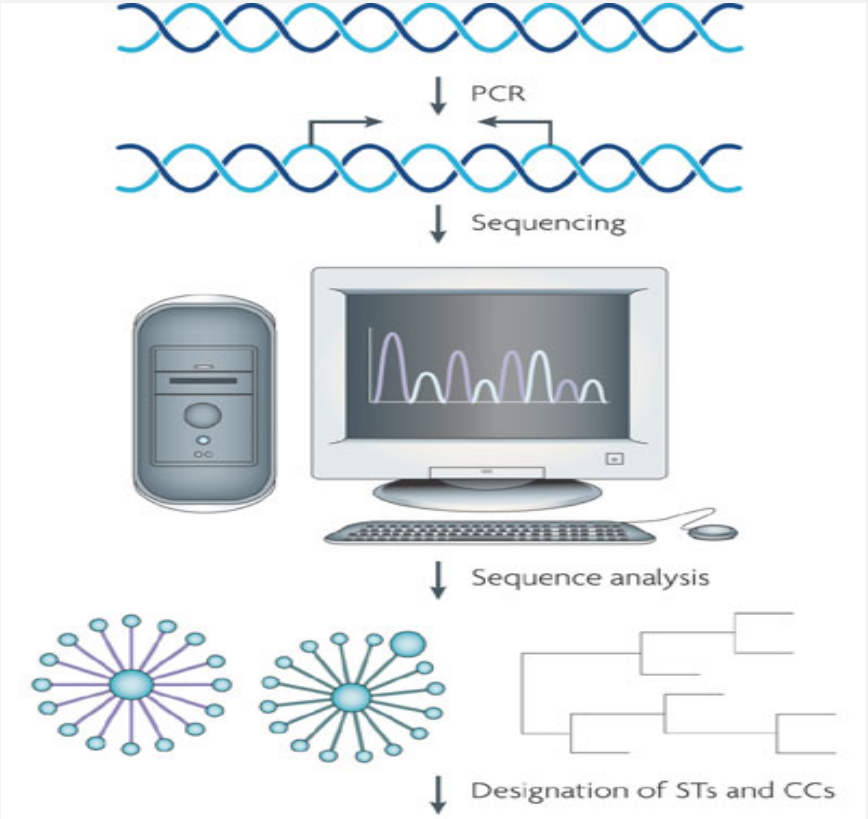
Gene	Primer sequence
fusA	TCTGGCSCARGARGAYCC
	GCCTCTTCGTARTGGTCRAA
lacF	GCTSTTCTGGAAGTCSCTST
	SAGRTTCCACCACTTGAAGC
gapA	GGCAATCAAGGTTGGYATCAACG
	ATCTCCAGGCACTTGTTSGARTAG
gltA	ATCTTGATCAGGTCACGCTCAAC
	AGCATCTTCAGCACGGCTTCGTT
gyrB	AAGTTCGACGACAACAGCTACAA
	GAMAGCACYGCGATCATGCCTTC
lepA	AAGCSCAGGTGCTCGACTCCAAC
	CGTTCCTGCACGATTTCCATGTG

Degenerate primers:

A mix of oligonucleotide sequences in which some positions contain a number of possible bases, giving a population of primers with similar sequences that cover all possible nucleotide combinations for a given protein sequence.

Degenerate base designation	Actual bases coded
N	A or C or G or T
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
K	G or T
M	A or C
R	A or G
S	C or G
W	A or T
Y	C or T

Creating MLST fingerprinting database



Chromosomal genes and allelic designation							ST	CC
arcC	*aroE*	*glpF*	*gmk*	*pta*	*tpiA*	*yqiL*		
1	1	1	1	1	1	1	1	1
1	4	1	4	12	1	10	5	5
3	3	1	1	4	4	3	8	8
3	3	1	1	4	4	16	250	8
3	3	1	12	4	4	16	247	8

Strain, species, or group[a]	Allele or sequence type[b]					
	lacF	*lepA*	*gyrB*	*fusA*	*gapA*	*gltA*
Xanthomonas sp. strain ETH12	1	1	1	1	1	1
X. perforans						
Group 1 (33)	2	2	2	2	2	2
Group 2 (14)	2	2	3	2	3	2
Xp4-20	2	3	3	2	2	2
X. euvesicatoria						
Group 1 (55)	3	4	3	3	3	3
Group 2 (8)	3	4	3	4	4	3
1605	3	5	3	3	3	3
Atypical Nigerian strains (4)	3	6	2	2	5	3
X. vesicatoria						
Group 1 (3)	4	7	4	5	6	4
Group 2 (3)	4	7	5	5	7	4
Group 3 (3)	4	7	5	6	8	5
X. gardneri						
Miscellaneous strains (28)	5	8	6	7	9	6
ICMP7383	6	8	7	8	10	6

What do we learn from MLST/A?

- Taxonomy: Identification and classification of bacteria. Interspecific and Intraspecific levels.
- Population genetics (High resolution population studies)
- Molecular evolution
- Epidemiology: Fingerprint database for identification and tracking of strains, track the source of disease outbreaks

MLST/A investigates whether recombination or mutation has played a major role in determining genetic variability. Different kinds of strain evolution are studied (diversifying, directional or purifying selections) as a prerequisite for understanding epidemiology and disease cycle of pathogen.

As a rule of thumb: Isolates with a 99% to 100% match can be considered to be very closely related to your isolate and may belong to the same pathovar or the same clonal line. Isolates that are between 96% and 99% identical to your isolate are probably of the same subspecies or genomospecies. Isolates with less than 96% identity are only distantly related to your isolate.

http://genome.ppws.vt.edu/cgi-bin/MLST/home.pl

Plant Associated and Environmental Microbes Database (PAMDB)

Home | Search | View Alleles | BLAST | Add/Edit Info | Help | Contact | Log out | VT

Logged in as sujan.timilsina@gmail.com working on *Pseudomonas*

PAMDB.org

Hosted at PPWS

We recommend these browsers

Compatability issues

Welcome to PAMDB.org!

Plant Associated and Environmental Microbes Database (PAMDB) is a multilocus sequence typing and analysis (MLST/MLSA) website and database specifically designed for identification of plant associated and environmental microbes and for the study of their epidemiology, population genetics, and molecular evolution. At this time, *Pseudomonas syringae*, *Xanthomonas* sp., *Ralstonia solanacearum*, *Acidovorax*, and *Clavibacter michiganensis* are organisms that are supported. Contact us if you are interested in adding additional organisms.

The more users contributing to the database, the more useful it will become for everyone. Please read Help to learn how to submit your data to the database and before start using this website.

At PAMDB you can:

- Compare DNA sequences from your unknown microbial isolates against isolates in the database to determine their identity based on DNA sequence similarity;
- Find where in the world isolates similar to yours have been previously isolated;
- Download DNA sequences from isolates in the database for individual loci or for several concatenated loci;
- Add substrates to PAMDB (for example: plants, soil, water, snow samples) from which you are going to isolate microbes. Once you have isolated microbes from substrates, you can add these isolates to your substrates including sequence data (see below);
- Add your isolates to PAMDB including exact location, time and pictures of the plant or environment from which the isolates came from; and
  Add DNA sequence data to your isolates and then download your sequences together with the sequences of all (or selected) isolates in the database so you can place your isolates on a phylogenetic tree.

Sequence comparison

- Sequence comparison in database

  - PAMDB provides Multilocus sequence analysis/typing (MLSA/T) database for 5 different plant pathogenic bacteria:

  Pseudomonas, Xanthomonas, Acidovorax, Ralstonia, and *Clavibacter*

- Multiple MLSA/T schemes are publicly provided.

MLSA datasets

- Sample datasets from *Pseudomonas floridensis*

  - *gyrB, gltA, gap1,* and *rpoD* genes were initially sequenced from representative strains using primers published in Hwang et al. 2005.

- Single genes compared with PAMDB and NCBI database.

- Single gene sequences were individually aligned and concatenated for phylogenetic analysis.

MLSA datasets

- Construct a maximum likelihood phylogenetic tree using ‘Pseudomonas_concatenated.fas’ using the same steps as in 16S sequence analysis.

- ‘Loropetali_sample_data.meg’ includes prealigned MLSA data along with reference strains. Run a phylogenetic tree using the data.

Whole Genome Sequence Analyses

Average Nucleotide Identity

- Average nucleotide identity (ANI) is the measure of nucleotide similarity between the two genomes.
- 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.

Average nucleotide identity (ANI)

- Download the following genomes using the same link: https://github.com/sujan8765/NPDN/tree/master/Pseudomonas_sample_data
  - GEV388.fna
  - P_cichorii_JBC1.fna
  - P_syringae_DC3000.fna
  - P_viridiflava_CFBP_1590.fna

Average Nucleotide Identity

- Use the following link to calculate ANI: http://enve-omics.ce.gatech.edu/ani/

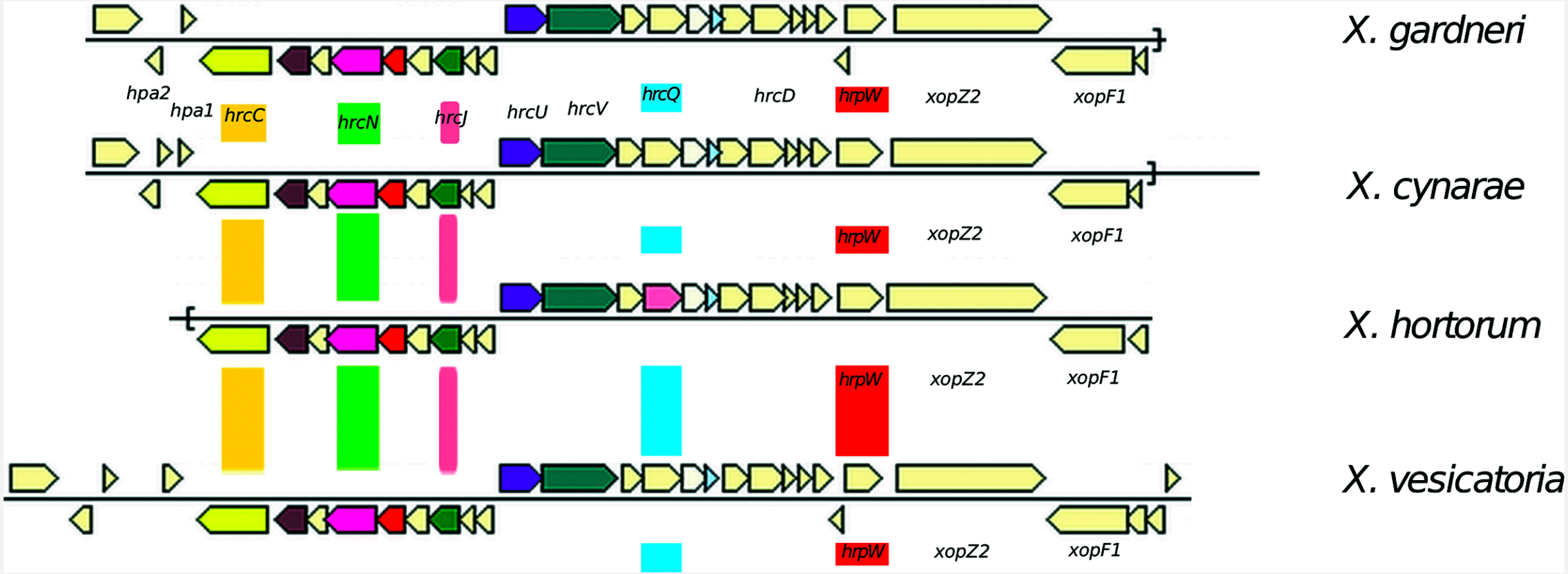
- Follow steps in the documentation to calculate pairwise average nucleotide identity among the strains.

InSilico DNA-DNA hybridization

- Genome-to-Genome Distance Calculator (GGDC) is commonly used along with ANI for species characterization with genome based data.
- Upload your query and reference genomes to http://ggdc.dsmz.de/ggdc.php.
- You will have to provide your email and will be notified once the analysis is complete.

Comparative Genomics

Comparative analyses to analyze the orientation and similarity of conserved genomic clusters.



Submitting sequences to NCBI

- NCBI GenBank:

  https://www.ncbi.nlm.nih.gov/WebSub/?tool=genbank

Real-Time PCR

- Real Time PCR: Detects the amount of DNA after each cycle with the use of fluorescent dye (SYBR Green) or TaqMan fluorescently labeled probes.

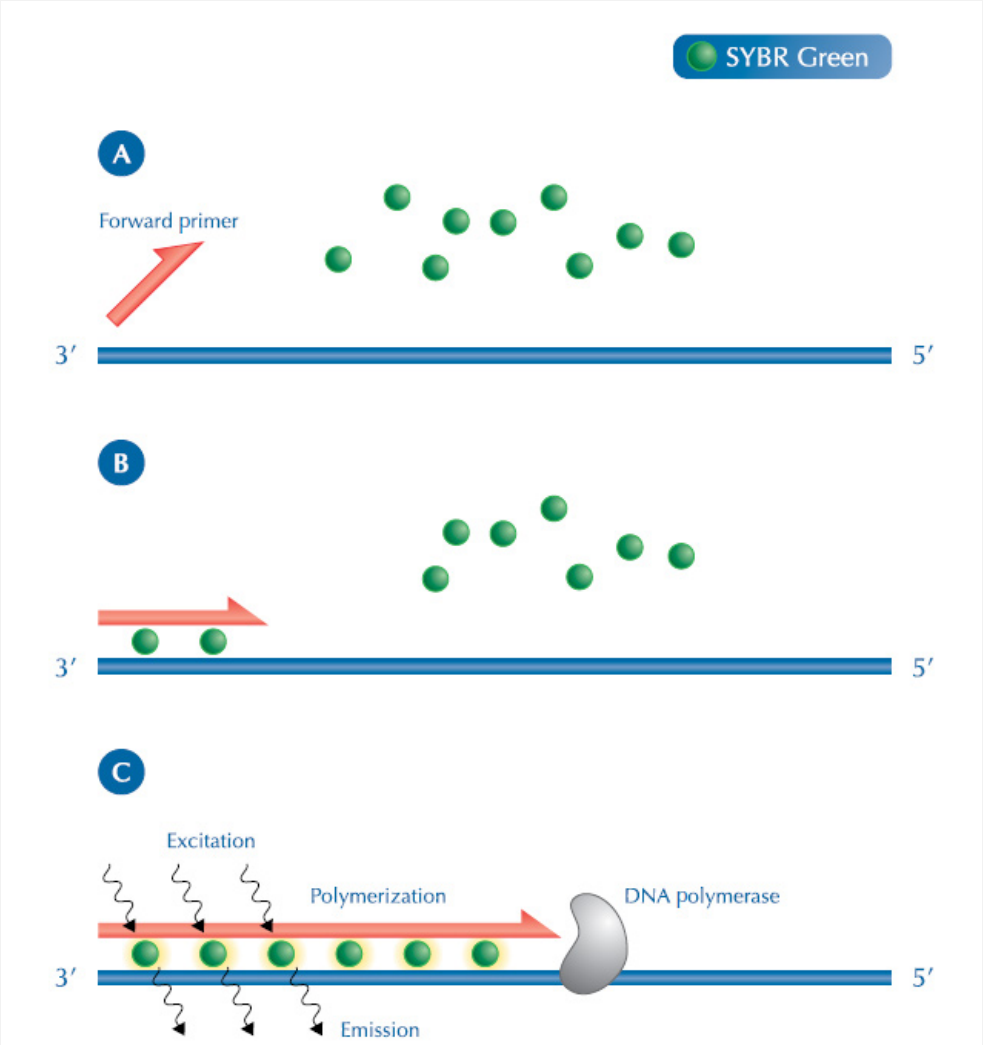
Table 2. Real-time TaqMan polymerase chain reaction probes (with the target single-nucleotide polymorphism in bold) and primers used in this study

Probe, primer[y]	Sequence	G+C (%)	T_m (°C)[z]
Xanthomonas perforans	5′/56-FAM/CGGGCAAGGAGC**C**ATCGCCTGT/31ABkFQ/-3′	68.1	66.6
X. euvesicatoria	5′/5TET/CGGGCAAGG**C**GCAATCGCCTGT/3BHQ_2/-3′	68.1	67.5
FP1	5′-CGTCGACGGCCTGGGCGA-3′	77.7	66.1
RP1	5′-CCGGTGCCTGCGCCTGGA-3′	77.7	66.7
X. gardneri	5′-/5TexRd-XN/TGCGCCAGCG**T**GACGGCACGC/3IAbRQSp/-3′	76.2	70.4
X. vesicatoria 1	5′-/5Cy5/TGCGCCAGCG**C**GATGGCACGC/3IAbRQSp/-3′	76.2	70.8
FP2	5′-AGGTCAGCCTGGGCGAGGT-3′	68.4	64.0
RP2	5′-TGAAGCCCACCACCTCGGC-3′	68.4	63.0

[y] Forward primer (FP) 1 and reverse primer (RP) 1 are specific for *X. perforans* and *X. euvesicatoria*, and FP2 and RP2 are specific for *X. vesicatoria* and *X. gardneri*.
[z] Melting temperature.

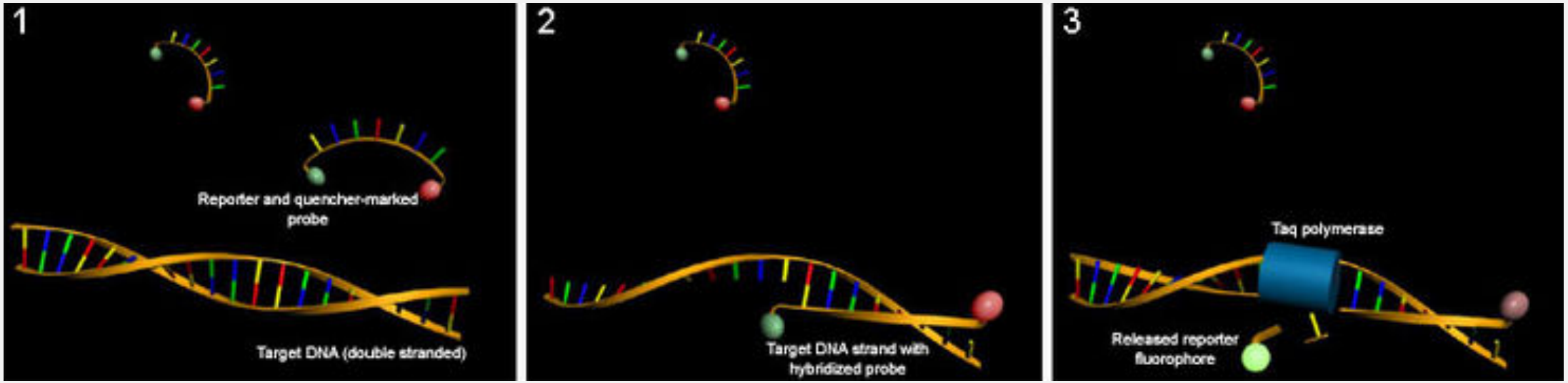
SYBR® Green based detection

- Is a Intercalating dye that emits fluorescence when bound to dsDNA
- It binds un-specifically to any dsDNA present (even unwanted primer-dimer products).
- Sensitivity and specificity are determined by the primers



TaqMan® based detection

- Probes are labeled with a fluorescent reporter dye at the 5‘end and a quencher dye at the 3' end
- A TaqMan® probe only emits fluorescence when it is bound to the target (2) and is cleaved by the *Taq* polymerase (3)



Real-Time PCR: Primer Design

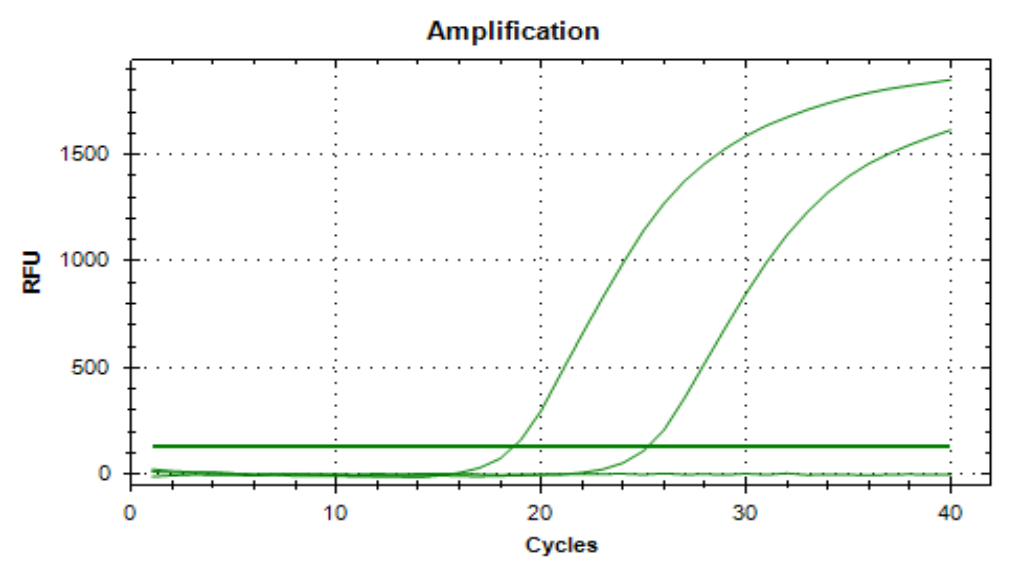
- Primer length should be 15-25 base pairs
- GC content can vary between 20-80%
- T_m of primers should match closely (usually 58-60°C)
- Avoid secondary structures and complementary pairing between forward and reverse primers
  - Avoid runs of an identical nucleotide.
- If designing TaqMan® assay,
  - Design the primers after you design the probe.
  - Design the primers as close to the probe as possible without overlapping.

Real-Time PCR: Probe Design

- No longer than 30 bp in length
- Keep the G-C content in the 20 to 80% range.
- Avoid runs of an identical nucleotide.
  - This is especially true for guanine, where runs of four or more Gs should be avoided.
  - Can affect hybridization efficiency due to secondary structure
- **Do not put a G on 5' end of the probe.**
  - Guanine can quench fluorescence
- The T_m should be 10°C higher than T_m of primers (usually 68-70°C)

Real-Time PCR: What is a Ct Value?

- The Ct (cycle threshold): the number of cycles required for the fluorescent signal to cross the threshold.
  - They are inversely proportional to the amount of target nucleic acid in the sample (i.e. the lower the Ct level the greater the amount of target nucleic acid in the sample)



Recombinase polymerase amplification

- Recombinase polymerase amplification (RPA)
  - Developed by TwistDX Ltd (Cambridge, UK)
- An isothermal, nucleic acid amplification method that utilizes three core enzymes
  - **Recombinase:** facilitates the binding of oligonucleotide primers to the targeted DNA template
  - **Single-stranded binding protein (SSB):** bind to the displaced strands of DNA and prevents the displacement of primers
  - **Strand-displacing polymerase:** recognizes the bound primer-recombinase complex and initiates DNA synthesis
- **Advantages:**
  - Short incubation times (10 to 30 min)
  - Operates at a low, single incubation temperature (37 to 42°C)
  - Maintains activity in PCR-inhibiting environments

RPA

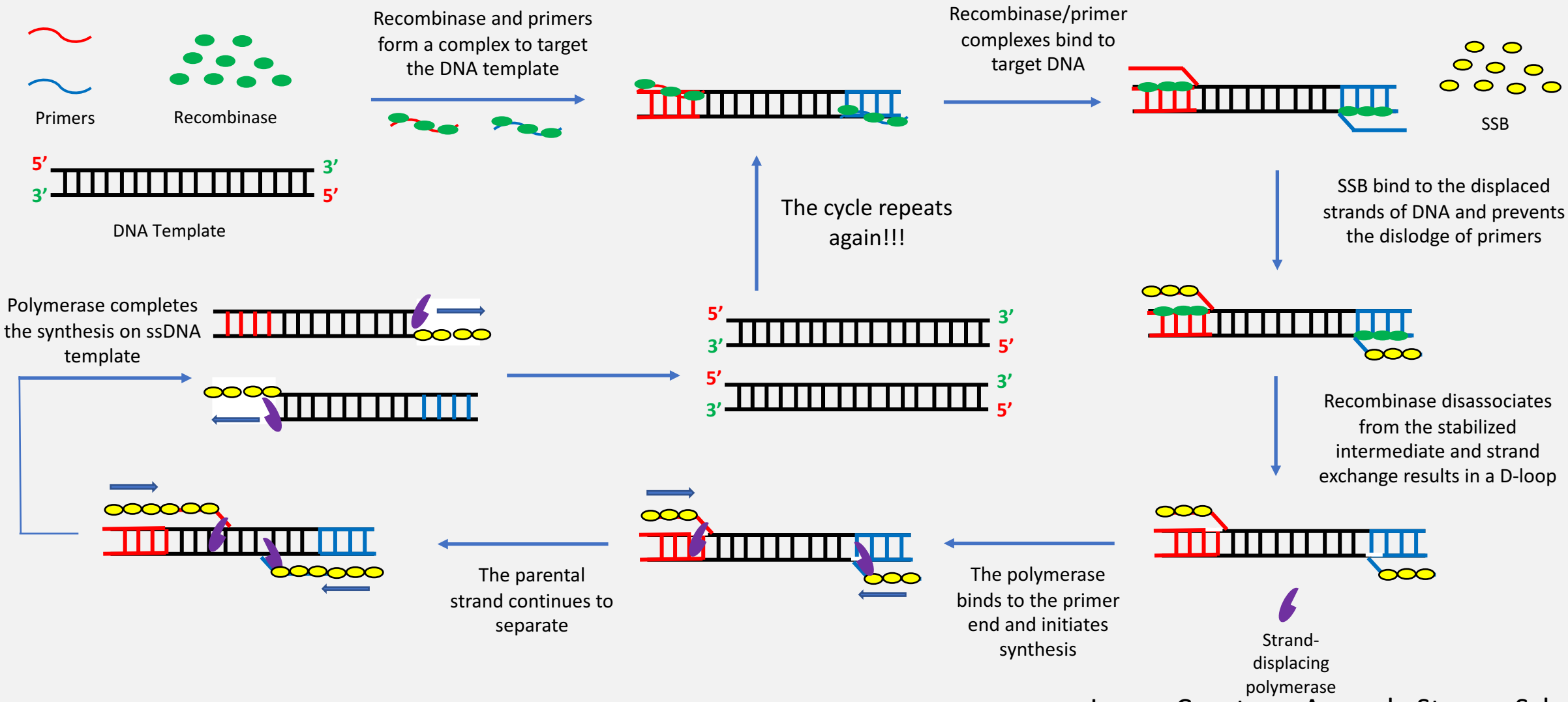
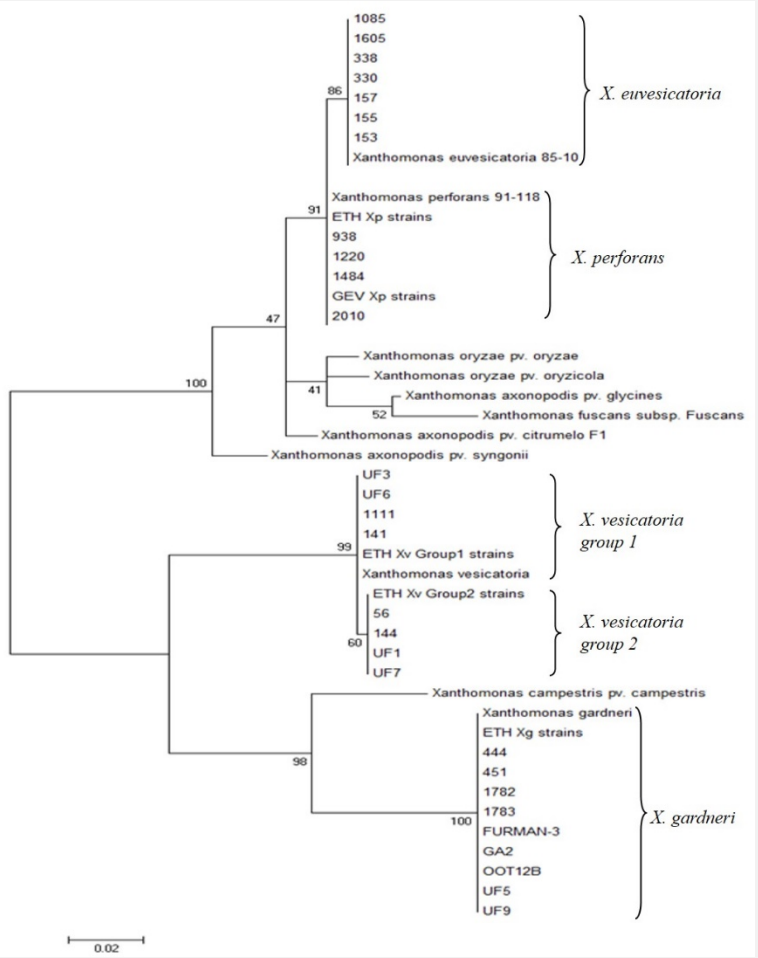


Image Courtesy: Amanda Strayer-Scherer

RPA

- RPA Exo-probes based on *hrpB* gene sequences to differentiate the four *Xanthomonas* species that cause bacterial spot of tomato
- Optimize RPA assays using Axxin T8-ISO instrument (TwistDx Ltd, Cambridge, UK) for both laboratory and field-based applications

	Probe Specificity			
Target Species	*X. euvesicatoria*	*X. perforans*	*X. gardneri*	*X. vesicatoria*
X. euvesicatoria	+	+	-	-
X. perforans	-	+	-	-
X. gardneri	-	-	+	-
X. vesicatoria	ND	ND	ND	ND



Condensed maximum likelihood phylogenetic tree with bootstrap values based on the *hrpB7* (Strayer et al. 2016).

Suggested Readings

- *Pseudomonas floridensis* sp. *nov*., a bacterial pathogen isolated from tomato. doi:10.1099/ijsem.0.002445
- 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. doi: 10.1099/ijsem.0.003104
- A multiplex real-time PCR assay differentiates four *Xanthomonas* species associated with bacterial spot of tomato. Doi: 10.1094/PDIS-09-15-1085-RE
- Bacterial Gall of Loropetalum chinense caused by *Pseudomonas amygdali* pv. *loropetali* pv. nov. doi: 10.1094/PDIS-04-17-0505-RE

Questions???

SUJAN TIMILSINA, Ph.D.

Department of Plant Pathology

University of Florida

Email: sujan.timilsina@ufl.edu

Twitter: @sujantimilsina

Web: https://sites.google.com/site/sujantimilsina