SUMMARY

Advent of genomics era is transforming the field of bacteriology. Now it is possible to sequence large number of bacterial strains in short time and affordable cost. The number of publically available genome sequences is exponentially increasing day by day. In last two decades, the number of genome sequence of bacterial strains has gone to nearly one lakh. Typing and analyzing bacterial population from genera to isolate level now becoming easier than ever. New genome based criteria like ANI and dDDH have emerged to establish identity of species that are replacing conventional methods of taxonomic studies. Whole genome information is vast and complex is used in total for understanding relationship of bacteria. In depth genome based studies have led to identification of set of housekeeping genes that can be used as reference markers in phylogenetic studies.

Pseudomonas syringae is a top pathogen among bacterial phytopathogen as reviewed by Mansfield and studied extensively for its pathogenicity (Mansfield et al.,2012). It is a serious pathogen of economically important plants like Tomato, beans, kiwi fruit, etc. It is also model plant pathogen. It has also life style of being epiphytic as well as pathogenic with plant hosts. Due to its pathogenic properties towards plants and host specificity it has been classified and named mainly as different pathovars. Since its first taxonomic classification in 1902 by van hall it was names as *Pseudomonas syringae*. However due to lack of proper identification system in case of *P. syringae* strains several isolates had been misclassified in previous studies and hence needs to be attended. It is only plant pathogen with rich genomic resources not only in terms of number of strains sequenced within this species complex. In addition, this also includes all reference pathovar strains sequenced that became available in NCBI GenBank. Bacterial strains belonging to same genus can display high amount variation in their genomes and much of it has originated from horizontal gene transfer events at multiple genomic sites. Hence the challenge is to study the relationship of large number of genome sequence and use such understanding to make sense of hyper-variable regions. This is particularly relevant to bacterial species like *P. syringae* that have pathogenic and diverse life styles. Hence systematic genome based studies on diversity is need of the hour for their management and finding evolution of genes/genomic regions.

In chapter 2, I have provided proof of concept to use of complete *rpoB* gene sequence as an ideal marker gene for taxonomic and phylogenetic study. Not only it is larger in size but its sequence similarities cutoff of 97.8% between isolates has been correlated with DDH, the gold standard of bacterial taxonomy. Further *rpoB* is also one of the reference marker genes identified from new genome based studies as relatively immune to horizontal gene transfer. My study revealed six potential species group (RSG) in 56 reference pathovar strains of *P. syringae* complex and a robust tree revealed accurate relationship among these RSGs. Using multilocus sequence data, I found that the RSGs in *P. syringae* complex have varying rate of mutation and recombination. This is also reflected in relatively higher nucleotide diversity in RSG2. Interestingly, modern genome based criteria like ANI and dDDH along with traditional MLST scheme failed to recognize RSG2. This may be because of pathogenic lifestyle of *P. syringae* species where high diversity is fuel for their success. Interestingly, I also discovered that RSG group has higher number of Non-Ribosomal Peptide Syntetases (NRPS) in their genome. It is possible that NRPS are providing advantage and ability to evolve rapidly as destructive pathogens of plants.

In addition, we also found half to quarter percentage of total number of genes encoded in the genome with altered GC content suggestive of rampant horizontal HGT events in RSGs which is also reflected in distribution of 11 species groups based on 70% cutoff for digital DNA-DNA hybridization score in 56 reference pathovars. However MLST and ANI based phylogroups were showing 7 species groupings. Hence it is not surprising that dDDH performed poorly in comparison to MLST and ANI. Ease and speed of using *rpoB* gene as phylogenomic marker allowed me to carrying out diversity of nearly 300 strains classified as *P. syringae* and its relatives of diverse ecological origin. This study allowed me to discover at least thirteen species in *P. syringae* complex and their relationship in a robust phylogenetic tree. Apart from identifying *P.syringae* complex, I could also clearly identify outgroup species such as *P. lutea*, *P. marginalis* and *P. abietaniphila* of this complex which is critical for comparative and evolutionary studies. *rpoB* based taxonomic and phylogenetic analysis also allowed me to understand several species of *Pseudomonas* with *P. syringae*. For example *P. fluorescens* and *P. putida* are entirely different species group however, *P. fluorescens* are more closely related to *P. syringae*. Traditionally, the *Pseudomonas* and *Xanthomonas* relationship is controversial since start of last century. In this regard, I used *rpoB* gene sequence of one of the oldest pathovar of *Xanthomonas* and newest pathovar for comparison. Using *rpoB* cutoff of 85.5% for genus

delineation, I also provided conclusive evidence that *Pseudomonas* and *Xanthomonas* are indeed separate genera.

In chapter 3, I carried out comparative genomics study of *P. syringae* and its relatives based on clear findings on phylogeny from chapter 2. Instead selecting random variations, I carried out studies on identification and sequence characterisation of a particular locus that is known to be hyper-variable in both size and sequence and packaged between conserved genes within these species. Lipopolysaccharide acts like a PAMP in bacteria, target of bacteriophages and also provides resistance to antimicrobial substances. Hence the gene clusters of LPS are highly variable and need to be systematically studied, identified for its variation at LPS biosynthetic locus in *P. syringae* complex. In this regard, my study led to submission of first genome sequence of type strain of two well-known pathogenic species of *Pseudomonas* i.e., *P. syringae* and *P. viridiflava*. My analysis led to identification of 23.8 kb LPS gene cluster in *P. syringae* reference strain that is located between *dnaJ-cheW/ychF-pth-rplY* locus. Interestingly, in *P. viridiflava* also, the LPS gene cluster is located at the same genomic location. However in *P. viridiflava*, the size is cluster is only 4.9 Kb but also less number of genes are present. Interestingly, a tRNA methionine gene is located besides dnaJ end of LPS cluster, flanking this variable locus and suggesting role of HGT in its ultravariation.

Initially I carried out variation studies at LPS locus in all the strains of *P. syringae and P. viridiflava* whose genome is available in NCBI. During the course of study, I carried out sequencing, assembly, annotation and submission of type strain of *P. syringae* pv. syringae and *P. syringae* pv. viridiflava in NCBI GenBank. Since I was getting complete gene cluster, I focused on draft genome of *P. syringae* and *P. viridiflava* strains where I could get complete gene cluster. I also included strains that are not only isolated as plant pathogens but also from diverse origin like soil, epilithon and rivers etc.

In depth comparison of coding regions and gene cluster led to me to identify four types/groups of LPS cassettes in *P. syringae* complex. To this, LPS cluster type I and III are predominantly present in most of the strains while group II and IV constitutes small part around 6% of total population diversity. The regions coding within the LPS clusters contains several important genes for LPS biosynthesis, transport and modification revealed after detailed annotation.These clusters describes itself as pathogenicity islands as it contains a conserved tRNA-met gene

sequence throughout every cluster at start. It also shows remarkable GC content differences compared with genomic GC content. These clusters are present between conserved *dnaJ-cheW/ychF-pth-rplY* locus. Several IS elements were also present in these LPS clusters suggesting dynamic nature of this locus. However, some strains mostly related to *P. viridiflava* and out group members had less coding sequences within this locus, missing some important genes like *wzm-wzt* transporters. Due to unique homology of *wzt* protein it has been used as LPS group markers as well. The relatively low GC content of the cassettes suggests their origin from horizontal gene transfer events. This suggest that the locus in under selection in *P. syringae* complex at species and strain level. The evolutionary drive in these strains were also reflected in findings of other significant coding regions such as bacteriocin, NRPS, phase coding and CRISPR sequences. RSG 2 member's shows significantly high numbers of NRPS coding clusters, adding to more genomic diversity in the species group compared to other RSGs.

Chapter 4 mainly describes the LPS clusters present in 62 reference *P. syringae* and *P. viridiflava* pathovars and their diversification in terms of different LPS groups/types as well as cassettes variations. An important gene i.e. *wzt* has been useful in determining LPS type groups based on its sequence homology differences and phylogenetic groupings. This study also reflects the findings of chapter three and reconfirms the facts about presence of heterogenic LPS groups and distribution in host specific plants with varied geographic locations.

Lack of some important metabolic genes in *P. viridiflava* and other phylogenetic near by strains in *P. syringe* species complex led us to discover a new alternative locus for LPS biosynthesis in these strains. Chapter 5 describes the study of this alternate locus of *cheW/ dctP-dctQ-dctM* for complementing the functional aspects of LPS biosynthesis in these strains and probably co-evolution of this locus has role in origin of *dnaJ-cheW/ychF-pth-rplY* locus. Further, we ascertained the specificity and exclusive site of this LPS locus with in genus *Pseudomonas* supported by fact that other *Pseudomonas* species do have the conserved lps site but devoid of LPS coding genes. *P. syringae* and *P. viridiflava* also lack any LPS coding region corresponding to *P. aeruginosa* B-band LPS locus, which affirms our findings.

Thus, the whole study focuses on the phylogenetic speciation with in *P. syringae* species complex and their functional groupings based on a robust phylogenomic marker *rpoB* gene. At the genomic level, a hypervariable region coding for LPS was discovered and studied in details

about their heterogeneity, diversity and evolution of LPS Pathogenicity Island in *P. syringae* and its relatives. This information will augment the understanding of host pathogen interactions of *P. syringae* and plants.