Comparative Genomics Analyses of Selected Microbes with Open Pan-genomes Highlight Their Evolutionary Dynamics at Functional Level

A Thesis

to be submitted by

Vikas Sharma (Roll No D12090)

for the award of the degree

of

Doctor of Philosophy

School of Basic Sciences

Indian Institute of Technology Mandi

Kamand, Himachal Pradesh-175005, India

February, 2019

Indian Institute of Technology Mandi
Mandi-175001, Himachal Pradesh, India

भारतीय प्रौद्योगिकी संस्थान मण्डी
मण्डी–175001, हिमाचल प्रदेश, भारत

Declaration by the Scholar

I hereby declare that the entire work embodied in this Thesis is the result of investigations carried out by me in the **School of Basic Sciences**, Indian Institute of Technology Mandi, India, under the supervision of **Dr. Tulika Prakash Srivastava**, and that it has not been submitted elsewhere for any degree or diploma. In keeping with the general practice, due acknowledgements have been made wherever the work described is based on finding of other investigators.

Place: Signature

Date: Name: **Vikas Sharma**

Indian Institute of Technology Mandi
Mandi-175001, Himachal Pradesh, India

भारतीय प्रौद्योगिकी संस्थान मण्डी
मण्डी–175001, हिमाचल प्रदेश, भारत

Declaration by the Research Advisor

I hereby certify that the entire work in this Thesis has been carried out by **Vikas Sharma** under my supervision in **School of Basic Sciences**, Indian Institute of Technology Mandi, and that no part of it has been submitted elsewhere for any Degree or Diploma.

Signature:
Name of the Guide: Dr. Tulika Prakash Srivastava
Date:

Dedicated to My Mother

Acknowledgement

First of all, praises and thanks to the Almighty God, for showering His blessings for the accomplishment of my research work.

This thesis is no exception, I pay sincere and special appreciations to several individuals. Many minds have indirectly or directly contributed to the research work presented in this thesis.

I feel the honour to express my sincere gratitude and warm regards to my advisor Dr. Tulika Prakash Srivastava. This work would not have been possible without her constant support, valuable guidance, encouragement on several occasions, numerous exciting discussions related to this research. Under her guidance, I successfully overcame many difficulties and learned a lot. I appreciate all the contributions of time, ideas and funding to make my PhD experience productive and stimulating. She always knew where to look for the answers to obstacles while leading me to the right source, theory and perspective.

I am grateful to Prof. T. A. Gonsalves, Director, IIT Mandi, for his help and support during my tenure in the institute. I thank my thesis doctoral committee members, Dr. Venkata Krishnan, Dr. Shyam Kumar Masakapalli, Dr. Rajanish Giri, and Dr. Varun Dutt. Their valuable advice, constructive criticism and extensive discussions about my work were a great help. I am also thankful to the office staffs Mr. Sumeet, Mr. Sushil Kumar Pal, Ms. Stuti, Ms. Sushma, Mr. Kuldeep, Mr. Monu, Mr. Prakash, Mr. Manoj, Mr. Suneel, Mr. Pawan, Mr. Anoop and Mr. Prateek for their support.

I take this opportunity to sincerely acknowledge, Indian Institute of Technology (IIT), Mandi, for providing me with the financial assistance during my Ph.D. tenure. I also thank Prof. Dr. Dieter Jendrossek (Universität Stuttgart) and Prof. Dr. Georg Sprenger (Universität Stuttgart), and Prof. Dr. Tilman Lamparter (Karlsruher Institut für Technologie) for the collaborative work.

I am thankful to my lab-mates for providing a stimulating and fun-filled environment. I am ever grateful to Dr. Shrawan Kumar and Fauzul Mobeen for his cooperation and fruitful discussions at several steps of research. I am also thankful to Pankaj, Mohit, Gaurav, Adil, Imran, Rohit, Suraj, Salam, Poonam, Prakash, Amar, and Abhishek for their care and moral support. They have helped me to survive all the stress and not letting me giveup. I am also grateful to my childhood friends, Akhilesh and Abhay for their constant support during my research work.

Finally, I would like to thank my beloved family members, for their unconditional love and support during my good and bad times. I couldn’t have completed this work without their consistent moral support.

Table of Contents

Chapter 1: General Introduction 1-30

1.1. Microbial genome evolution 2-3

1.2. Pan-genome 3-7

1.2.1. History 3-5

1.2.2. Core and dispensable genes 5-6

1.2.3. Open and closed pan-genomes 6-7

1.3. Mechanisms of genome expansion 7-12

1.3.1. Horizontal gene transfer 8-9

1.3.2. Gene duplication 10

1.3.3. Selective pressure 11-12

1.4. Approaches to study microbial genomes 12-17

1.4.1. Comparative genomics 12-16

1.4.1.1. Multiple strain sequencing 13-14

1.4.1.2. Multiple genomes of a given taxon in a metagenomic survey 14-16

1.4.2. Hybridization-based techniques 16-17

1.5. Motivation 17-19

1.6. Aim of thesis 20-22

1.7. Overview of the thesis chapters 23-24

1.8. References 24-30

Chapter 2: Materials and Methods 31-72

2. Materials and methods 32-72

2.1. Sequence analysis 32-35

2.1.1. BLASTP 32

2.1.2. BLASTN 32-33

2.1.3. RPS-BLAST 33

2.1.4. ClustalW 33

2.1.5. Clustal Omega 33-34

2.1.6. T-Coffee 34

2.1.7. PAL2NAL 34

2.1.8. MAUVE 34-35

2.1.9. DIAMOND 35

2.1.10. USEARCH 35

2.1.11. HMMER 35

2.2. Phylogenetic analysis 36-40

2.2.1. MEGA 36

2.2.2. MEGA-CC 36

2.2.3. Neighbor-Joining method 36

2.2.4. BIONJ algorithm 36-37
2.2.5. Maximum composite likelihood method 37
2.2.6. JTT matrix-based model 37
2.2.7. Poisson correction method 37
2.2.8. Maximum likelihood method 38
2.2.9. Kimura 2-parameter model 38
2.2.10. Bootstrap method 38
2.2.11. PhyML 39
2.2.12. UBCG 39
2.2.13. GSI 39
2.2.14. FigTree 39-40
2.3. Genome-based taxonomic analysis 40-43
2.3.1. ANI 40
2.3.2. OrthoANI 40
2.3.3. AAI 40-41
2.3.4. POCP 41
2.3.5. Bray-Curtis dissimilarity index 41-42
2.3.6. *In silico* DDH 42
2.3.7. Jspecies 42-43
2.3.8. GGDC 43
2.4. Pan-genome analysis 43-44
2.4.1. BPGA 43
2.4.2. GET_HOMOLOGUES 44
2.5. Selective pressure analysis 44-45
2.5.1. Kullback-Leibler divergence 44-45
2.5.2. KaKs_Calculator 45
2.5.3. NG Method 45
2.5.4. HyPhy 45
2.6. Horizontal gene transfer analysis 46
2.6.1. T-REX 46
2.6.2. Bipartition dissimilarity 46
2.7. Recombination analysis 46-50
2.7.1. Datamonkey 46
2.7.2. GARD 47
2.7.3. Pairwise homoplasy index test 47
2.7.4. SplitsTree 47
2.7.5. Tajima's D test 47-48
2.7.6. RDP software 48
2.7.7. 3Seq 48
2.7.8. Bootscan 48

2.7.9. Chimaera 48-49
2.7.10. GENECONV 49
2.7.11. MaxChi 49
2.7.12. RDP 49
2.7.13. SiScan 50
2.7.14. LDhat 50
2.7.15. DnaSP 50
2.7.16. Detection of gene conversion tracts 50
2.8. Automatic pipelines for complete genome annotation 51
2.8.1. RAST 51
2.8.2. IMG/ER 51
2.9. Functional analysis 51-60
2.9.1. NR database 51
2.9.2. UniProt 51-52
2.9.3. COGs 52
2.9.4. cdd2cog.pl 52-53
2.9.5. KEGG 53
2.9.6. BlastKOALA 53
2.9.7. WebMGA 53-54
2.9.8. eggNOG 54
2.9.9. Blast2GO 54
2.9.10. Interproscan 54
2.9.11. SecretomeP 54-55
2.9.12. TMHMM 55
2.9.13. TMMOD 55
2.9.14. LipoP 55-56
2.9.15. SignalP 56
2.9.16. CELLO 56
2.9.17. PSORTb 56
2.9.18. Inmembrane 57
2.9.19. TCDB 57
2.9.20. dbCAN 57
2.9.21. MvirDB 57-58
2.9.22. EffectiveDB 58
2.9.23. DEG 58
2.9.24. P2RP 58-59
2.9.25. MAHMI 59
2.9.26. InnateDB 59
2.9.27. IEDB 59-60
2.9.28. CARD 60

2.10. Protein-protein interaction analysis 60-61
2.10.1. BIPS 60-61
2.10.2. STRING 61
2.11. Regulatory element analysis 61
2.11.1. NNPP 61
2.11.2. CISTER 61
2.12. Mobilome analysis 62-65
2.12.1. IslandViewer 62
2.12.2. IslandPick 62
2.12.3. SIGI-HMM 62
2.12.4. IslandPath-DIMOB 63
2.12.5. Islander 63
2.12.6. ISfinder 63
2.12.7. ISsaga 63
2.12.8. PHAST 64
2.12.9. PHASTER 64
2.12.10. PHACTS 64
2.12.11. CRISPRfinder 65
2.13. Miscellaneous analysis 65-66
2.13.1. Unipro UGENE 65
2.13.2. NAViGaTOR 65
2.13.3. GENE-E 66
2.14. References 66-72
Chapter 3: Comparative Genomics and Pan-genome Analyses of the Family *Herpesviridae* **73-98**
3.1. Abstract 74
3.2. Introduction 74-77
3.3. Materials and methods 77-80
3.3.1. Materials 77
3.3.2. Methods 77-80
3.3.2.1. Pan-genome analysis 78
3.3.2.2. Core-genome analysis 78-79
3.3.2.3. Epitope prediction 79
3.3.2.4. Phylogenetic reconstruction 79
3.3.2.5. Cis-element prediction 79-80
3.4. Results and discussion 80-93
3.4.1. The viral family *Herpesviridae* has an open pan-genome 80-83
3.4.2. The essential core-genome of the viral family *Herpesviridae* has two genes 83
3.4.3. Phylogenetic analysis of the family *Herpesviridae* reveals sharp

splits of its subfamilies 83-85
3.4.4. Differences exist in the genomic attributes of human herpesviruses and non-human herpesviruses belonging to the viral family *Herpesviridae* 85-87
3.4.5. Products of core gene set of human herpesviruses harbor putative epitopes 87-90
3.4.6. Data mining of the genomes of human herpesviruses predicts cis-elements 90-93
3.5. Summary 94-95
3.6. References 95-98
Chapter 4: Comparative Genomics and Pan-genome Analyses of Bifidobacteria 99-166
4.1. Abstract 100-101
4.2. Introduction 101-105
4.3. Materials and methods 105-112
4.3.1. Materials 105-106
4.3.2. Methods 106-121
4.3.2.1. Pan-genome analyses 106
4.3.2.2. Phylogenetic analysis 106-108
4.3.2.3. Selection pressure analyses 108
4.3.2.4. Genome-size variation and mobilome analyses 108-109
4.3.2.5. Functional analyses 109-111
4.3.2.6. Interspecific Interactions between human and bifidobacteria 111-112
4.3.2.7. Interactome analysis of bifidobacterial core proteins 112
4.4. Results and discussion 112-156
4.4.1. Open pan-genome of the genus *Bifidobacterium* 112-114
4.4.2. Core-genome based phylogeny of the genus *Bifidobacterium* 115-117
4.4.3. Open Pan-genome of the probiotics and human-gut strains of the genus *Bifidobacterium* and its relation to selection pressure 117-125
4.4.4. Functional exploration of core-genome of the probiotic and human-gut strains of bifidobacteria 125-136
4.4.4.1. Evolutionary conserved basic features 125-128
4.4.4.2 Evolutionarily conserved probiotic traits 129-132
4.4.4.3. Evolutionarily conserved survival traits 132-136
4.4.5. Conserved protein-protein interactions of the human host and the probiotic and human-gut strains of bifidobacteria 136-141
4.4.6. Functional evolution versus genome-size variations among the probiotic and human-gut strains of bifidobacteria 141-156
4.5. Summary 156-157

4.6. Online resources 157-158

4.7. References 158-166

Chapter 5: Comparative Genomics and Pan-genome Analyses of Natural Rubber Degrading Bacteria 167-311

5.1. Abstract 168-169

5.2. Introduction 169-175

5.3. Materials and methods 176-193

5.3.1. Materials 176-177

5.3.1.1. Genome annotation and taxonomy analysis of *Xanthomonas* sp. 35Y 176

5.3.1.2. *In silico* functional and evolutionary analyses of rubber oxygenases (RoxA and RoxB) 176

5.3.1.3. Pan- and core-genome analyses of the order *Nevskiales* 176

5.3.1.4. Taxonomy analysis of the order *Nevskiales* 177

5.3.2. Methods 177-193

5.3.2.1. Genome annotation and taxonomy analysis of *Xanthomonas* sp. 35Y 177-182

5.3.2.1.1. Basic genome analysis 177-180

5.3.2.1.2. Genome mining of strain 35Y for the genes involved in degradation of natural rubber and biosynthesis of secondary metabolites 180

5.3.2.1.3. Taxonomy analysis of *Xanthomonas* sp. 35Y 180-182

5.3.2.2. Pan- and core-genome analyses of the order *Nevskiales* 182-183

5.3.2.3. *In silico* functional and evolutionary analyses of rubber oxygenases (RoxA and RoxB) 183-188

5.3.2.3.1. Assignment of orthologs 183

5.3.2.3.2. Mining of RoxA and RoxB orthologs 183

5.3.2.3.3. Sequence alignment 183-184

5.3.2.3.4. Phylogenetic tree reconstruction 185-186

5.3.2.3.5. Functional analyses 186

5.3.2.3.6. Horizontal gene transfer analysis 186

5.3.2.3.7. Intragenic recombination and gene conversion analyses 187

5.3.2.3.8. Selective pressure analysis 188

5.3.2.3.9. Genome mining for the prediction of natural rubber-degradation pathway 188

5.3.2.4. Taxonomy analysis of the order *Nevskiales* 190-193

5.3.2.4.1. *In silico* investigations 190-192

5.3.2.4.2. Experimental investigations 193

5.4. Results and discussion 193-298

5.4.1. *Xanthomonas* sp. 35Y belongs to a novel species of the genus

Steroidobacter 193-200
5.4.1.1. Description of *Steroidobacter cummioxidans* sp. nov. 200
5.4.2. Basic features of the draft genome of *S. cummioxidans* 35Y 200-201
5.4.3. Functional exploration of the genome of *S. cummioxidans* 35Y 201-237
5.4.3.1. The core-genome of *S. cummioxidans* 35Y is critical for its survival 201-208
5.4.3.2. Central energy metabolism 208-209
5.4.3.3. NR-degradation potential of *S. cummioxidans* 35Y 209-219
5.4.3.4. Biosynthesis of industrially relevant secondary metabolites 219-222
5.4.3.5. Pathogenesis 222-233
5.4.3.6. Additional genome and metabolic features 234-235
5.4.3.7. Mobile elements of *S. cummioxidans* 35Y 235-236
5.4.4. Functional and evolutionary exploration of Gram-negative natural rubber degrading bacteria 236-261
5.4.4.1. Prokaryotic distribution and diversity of rubber oxygenases (RoxA and RoxB) proteins 236
5.4.4.1.1. *In silico* mining and phylogenetic analysis of rubber oxygenases (RoxA and RoxB) 236-239
5.4.4.1.2. Functional analyses of rubber oxygenases (RoxA and RoxB) 239-244
5.4.4.1.3. Evolutionary analyses of *RoxA* and *RoxB* genes 244-261
5.4.4.2. Genome mining for the pathway of natural rubber-degradation 261-275
5.4.5. The order *Nevskiales* has an open pan-genome 275-277
5.4.6. Taxonomy re-assessment of *S. cummioxidans* 35Y and the order *Nevskiales* 278-298
5.4.6.1. Description of *Aurantibaculum* gen. nov. 296
5.4.6.2. Description of *Aurantibaculum cummioxidans* comb. nov. 296-297
5.4.6.3. Description of *Steroidobacteraceae* fam. nov. 297-298
5.5. Summary 298-302
5.6. Online resources 302-303
5.7. References 303-311
Chapter 6: Comprehensive Genomic Characterization and Pan-genome Analysis of *Phormidium lacuna* sp. nov., strain HE10JO 313-372
6.1. Abstract 314-315
6.2. Introduction 315-318
6.3. Materials and methods 318-323
6.3.1. Materials 318
6.3.2. Methods 318-323
6.3.2.1. Pan- and core-genome analyses 318-320

6.3.2.2. Genome annotation and functional analyses 320-321

6.3.2.3. Phylogenetic analyses 321-322

6.3.2.3.1. 16S rRNA gene-based analysis 321

6.3.2.3.2. Core-proteome based analysis 321

6.3.2.3.3. Average nucleotide identity and *in silico* DNA-DNA hybridization based analyses 322

6.3.2.3.4. Evolutionary analysis of nitrogenase 322

6.3.2.3.5. Evolutionary analysis of bidirectional hydrogenase 323

6.4. Results and discussion 323-365

6.4.1. The order *Oscillatoriales* has an open pan-genome 323-328

6.4.2. Strain HE10JO belongs to a novel species of the genus *Phormidium* 328-336

6.4.3. Basic genome features of *Phormidium lacuna* HE10JO 336-337

6.4.4. Transport and binding proteins of *Phormidium lacuna* HE10JO 337-341

6.4.5. Storage compounds of *Phormidium lacuna* HE10JO 341

6.4.6. Photosynthesis and other light effects of *Phormidium lacuna* HE10JO 341-348

6.4.7. Carbon catabolism of *Phormidium lacuna* HE10JO 348-350

6.4.8. Nitrogen metabolism and hydrogenases of *Phormidium lacuna* HE10JO 351-361

6.4.9. Other genes of *Phormidium lacuna* HE10JO of biotechnological importance 361-363

6.4.10. Other genome features of *Phormidium lacuna* HE10JO 363-365

6.5. Summary 365-366

6.6. Online resources 367

6.7. References 367-372

Chapter 7: Conclusions and Future Prospects **373-378**

7.1. Conclusions 374-376

7.2. Future prospects 376-378

Appendices **379-422**

Appendix A 380-418

Appendix B 419-422