

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
DEPARTAMENTO DE BIOLOGIA GERAL
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA



PhD Thesis

**Comparative Microbial Genomics: Pangenomics and
Pathogenomics of *Corynebacterium*, *Campylobacter* and
*Helicobacter***

Amjad Ali

Prof. Dr. Vasco Ariston de Carvalho Azevedo

BELO HORIZONTE

March - 2013

AMJAD ALI

**Comparative Microbial Genomics: Pangenomics and Pathogenomics
of *Corynebacterium*, *Campylobacter* and *Helicobacter***

Thesis presented as partial requirement
for the degree of Doctor of Philosophy in
Genetics, to the Department of General
Biology at the Institute of Biological
Sciences, Federal University of Minas
Gerais.

SUPERVISOR: Prof. Dr. Vasco Ariston de Carvalho Azevedo

Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Belo Horizonte – MG
March – 2013

Ali, Amjad

Microbial comparative genomics: pangenomics and pathogenomics of corynebacterium, campylobacter and Helicobacter. [manuscrito] / Amjad Ali. – 2013.

Orientador: Vasco Ariston de Carvalho Azevedo.

Tese (doutorado) – Universidade Federal de Minas Gerais,
Departamento de Biologia Geral.

1. Corynebacterium - Teses. 2. Pangênômica - Teses. 3. Patogênômica - Teses. 4. Genômica - Teses. 5. Genômica comparativa - Teses. 6. Campylobacter - Teses. 7. Helicobacter - Teses. 8. Candidatos vacinais. 9. Alvos de drogas. 10. Ilhas genômicas. 11. Genética - Teses. I. Azevedo, Vasco Ariston de Carvalho. II. Universidade Federal de Minas Gerais. Departamento de Biologia Geral. III. Título.

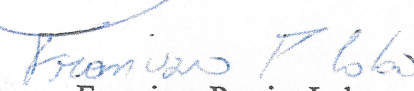
CDU: 576.85

**"Comparative Microbial Genomics: Pangenomics and Pathogenomics
of Corynebacterium , Campylobacter and Helicobacter "**

Amjad Ali


Tese aprovada pela banca examinadora constituída pelos Professores:


Vasco Ariston de Carvalho Azevedo - Orientador
UFMG


Francisco Pereira Lobo
EMBRAPA/CAMPINAS


José Miguel Ortega
UFMG


Liza Figueiredo Felicori Vilela
UFMG


Rommel Jucá Ramos
UFPA

Belo Horizonte, 27 de março de 2013.

I dedicate this work to my father (RIP), who has always been a source of inspiration for me. I also dedicate this efforts to those of my country fellows who are struggling for the betterment of the homeland.

ACKNOWLEDGEMENTS

All the praises be to Almighty Allah, Who is merciful, gracious, and whose bountiful blessings enabled me to persuade higher ideals of life. All regards and respects to the Holly Prophet Muhammad PBUH for guiding mankind towards the straight path of life. My sincere gratitude for my parents, my siblings and family for their support, love and friendship. I am very much grateful to my supervisor Prof. Dr. Vasco Azevedo for his continued technical and moral support, and guidance received. Thanks to Dr. Anderson Miyoshi for his support. I am thankful to Prof. David Ussery from center for biological sequence analysis, Technical University of Denmark for his full time availability and guidance. I acknowledge the faculty of Graduate Studies, the Coordinator, the Secretariat staff and fellow students for their cooperation and support. I am grateful to TWAS-CNPq for the financial assistance in the form of PhD Fellowship. My cordial thanks to all the colleagues in Laboratory of Cellular and Molecular Genetics (LGCM), for their pleasant company and guidance throughout my stay. I am very much thankful to my friends living all over the world for their moral support, whenever it was required to me.

My sincere thanks to everyone around!

CONTENTS

LIST OF ABBREVIATIONS	7
GLOSSARY	8
ACRONYMS.....	9
LIST OF FIGURES	11
LIST OF TABLES	11
Abstract	12
List of Papers	13
I. PRESENTATION	15
I.I Research Groups, Collaborators and Funding.....	16
I.II Structure of the Manuscript.....	19
II. INTRODUCTION	21
II.I Genome and Genomics	22
II.I.I Microbial Evolution and Diversity	23
II.I.II Characteristics of Microbial Genomes	25
II.I.III Levels of Microbial Genomics	26
II.I.IV Tree of Life and Statistics of Sequenced Genomes	27
II.II Microbial Comparative Genomics.....	29
II.II.I Gene/protein Analysis.....	31
II.II.II Intra-Genome Comparison (Single Genome Analysis).....	32
II.II.III Inter-Genome Comparison (Pairwise Alignment)	33
II.III Pangenomics	34
II.III.I Open and Close Pangenome	36
II.III.II Application of Pangenome in Vaccine Design	37
II.III.III Essential Genes and Minimal Genome	39
II.IV Pathogenomics.....	40
II.IV.I Genome Plasticity Analysis	41
II.IV.II Pathogenicity Islands	42
II.IV.III Virulence Factors in Pathogens	43
II.IV.IV Drug Targets Prediction	44
II.IV.V Vaccine Candidates Identification.....	46
II.IV.VI Reverse Vaccinology Approach.....	47
II.V Introduction to Bacterial Genera and Species Selected	48
II.VI Hypothesis	50

III. OBJECTIVES.....	52
III.I General Objectives	53
III.II Specific Objectives	53
IV. METHODOLOGY	54
IV.I General Methodology	55
IV.II FLOWCHART.....	56
CHAPTER 1	58
CORYNEBACTERIUM.....	58
1.1 Review Article	59
Microbial Comparative Genomics: An Overview of Tools and Insights into the Genus <i>Corynebacterium</i> . <i>J Bacteriol Parasitol</i> . March 2013.	
1.2 Conclusion	76
CHAPTER 2	77
CAMPYLOBACTER.....	77
2.1 Research Article	78
Campylobacter fetus subspecies: Comparative genomics and prediction of potential virulence targets. <i>GENE</i> 2012.	
2.2 Conclusion	91
CHAPTER 3	92
HELICOBACTER	92
3.1 Research Article	93
Computational comparative genomic based insights into human gastric pathogen <i>Helicobacter pylori</i> (38 species), essential features and species pangenome. (Manuscript).	
3.2 Conclusion	125
V. DISCUSSION	126
VI. CONCLUSION AND FUTURE PERSPECTIVES	151
VII. BIBLIOGRAPHY.....	154
VIII. CURRICULUM VITAE	167
IX. APPENDIX.....	176
IX.I Chapter 1. <i>Corynebacterium</i>	177
IX.II Chapter 2. <i>Campylobacter</i>	179
IX.III Chapter 3. <i>Helicobacter</i>	182

LIST OF ABBREVIATIONS

INSTITUTES, DATABANKS AND FUNDING AGENCIES

CNPq: National Council for Scientific and Technological Development

DDBJ: DNA Databank of Japan

EMBL: European Molecular Biology Laboratory

FAPESPA: Fundação Amazonia Paraense

GOLD: Genomes OnLine Database

ICB: Instituto de Ciências Biológicas

LGCM: Laboratory of Cellular and Molecular Genetics

LPDNA: Laboratory of DNA Polymorphism

NCBI: National Center for Biotechnology Information

NIH: National Institutes of Health

RGMG: Minas Gerais Genome Network

RPGP: Network Paraense Genomics and Proteomics

TIGR: The Institute For Genomic Research

TWAS: The Academy of Sciences for the Developing World

UFMG: Federal University of Minas Gerais

UFPA: Federal University of Pará

GLOSSARY

Core genome: set of genes commonly shared by all genomes in a species or taxa.

Macroevolution: any evolutionary change above the level of species.

Microevolution: any evolutionary change under the level of species.

ORFan (orphan): gene for which no homolog is found in current databases.

Orthologs: genes with a relationship that arose from a speciation event.

Pan genome: the union of all the genes that can be found in a species.

Paralogy: relationships between two duplicated genes.

ACRONYMS

ACT: Artemis Comparison Tool

BLAST: Basic Local Alignment Search Tool

BLASTn: Basic Local Alignment Search Tool (nucleotide)

BLASTp: Basic Local Alignment Search Tool (protein)

CDS: Coding sequence

Cff: Campylobacter fetus subspecies fetus 82-40

Cfv: Campylobacter fetus subspecies venerealis NCTC 10354^T

CGFs: Core gene families

CLA: Caseous Lymphadenitis

CMN: Corynebacterium, Mycobacterium and Nocardia

COG: Cluster of Orthologous Genes

Cp: *Corynebacterium pseudotuberculosis*

CRISPR: Clustered Regularly Interspaced Short Palindrome Repeats

DNA: Deoxyribonucleic acid

EGFs: Essential gene families

G+C: Guanine+Cytosine

GFs: Gene families

GO: Gene Ontology

GSS: Genomic Survey Sequence

HGT: Horizontal gene transfer

kb: kilobase (thousand bases)

Mb: mega bases (million bases)

ORF: Open Reading Frame

PAIs: Pathogenicity islands

pb: base pair

PFGE: Pulsed- Field Gel Electrophoresis

PGFs: Pan gene families

PLD: phospholipase D

rDNA: Ribosomal DNA

RFLP: Restriction Fragment Length Polymorphism

RNA: Ribonucleic acid

rRNA: ribosomal RNA

SOLiD: Sequencing by Oligonucleotide Ligation and Detection

T4SS: Type IV secretion systems;

tBLASTx: Basic Local Alignment Search Tool (Translated nucleotides)

tRNA: transfer RNA

LIST OF FIGURES

Figure 1.28
Tree of life and phylogenetic distribution of sequenced organisms.

Figure 2.36
Illustration of the genes distribution in the pangenome of bacteria.

Figure 3.43
Graphical illustration of a model pathogenicity island.

Figure 4.57
A schematic representation of the methodology in flowchart.

LIST OF TABLES

Table 1.40
Representative bacterial species with estimated essential genes.

Abstract

In the last decade, robust sequencing technologies have revolutionized the genomic science. As a result, comparative genomics is now recognized as a new discipline. Comparative microbial genomics exploits both similarities and differences in the genomes, proteome, transcriptome, and regulatory regions of different organisms to infer the evolutionary relations, along with conserved and unique characteristics of species. These analyses have resulted in some surprising biological discoveries in the recent past. This study presents comparative genomic analysis of multiple pathogenic and non-pathogenic bacteria from related species, to dissect the genomic information and to get insights into evolutionary relationships, conserved information and mechanisms of pathogenicity. Starting from genus *Corynebacterium*, 11 representative species are analysed and compared, resulting in 741 conserved **Gene Families** (GFs) in all of them, and significant intra-species proteome similarities (98-99%) were observed. Subsequently, the pan- (7059 GFs) and core genome (552 GFs) of genus *Campylobacter* is estimated. A detailed comparative pathogenomic study of *Campylobacter fetus* subspecies resulted in identification of common and novel regions associated with pathogenicity; and species specific virulence factors and vaccine candidates have been characterized. Furthermore, comparative genomics and pathogenomics analysis of the genus *Helicobacter* (46 genomes) is accomplished. 38 *Helicobacter pylori* were found to share 1,185 core gene families representing ~77% of the average genome size. The core **essential genes families** (EGFs) are ascertained, and explored for potential therapeutics against *H. pylori*. In conclusion, we propose that, these observed genomic variations, species specific features and core virulence factors will enhance understanding of the lifestyle of the organisms, and will contribute to the development of antibiotics, drugs and vaccines.

List of Papers

This thesis is presented in articles format and is divided into three chapters based on studied organisms and the following articles.

1. **Amjad Ali**, Siomar C Soares, Anderson R Santos, Eudes Barbosa, Debmalya Barh, Syeda M. Bakhtiar, Syed S. Hassan, Anderson Miyoshi, Vasco Azevedo, **Microbial Comparative Genomics: An Overview of Tools and Insights into the Genus *Corynebacterium***. *J Bacteriol Parasitol* 2013, 4:3.
2. **Amjad Ali**, Siomar C Soares, Anderson R Santos, Luis C Guimarães, Eudes Barbosa, Sintia S Almeida, Vinícius AC Abreu¹, Adriana, R Carneiro, Rommel TJ Ramos, Syeda M Bakhtiar, Syed S Hassan, David W Ussery, Stephen On, Artur Silva, Maria P Schneider, Andrey P Lage, Anderson Miyoshi, Vasco Azevedo. **Campylobacter fetus subspecies: Comparative Genomics and Prediction of Potential Virulence Targets**. *Gene* 2012 vol. 508 (2) p. 156-145;
3. **Amjad Ali**, Siomar C Soares, Syeda M Bakhtiar, Sandeep Tiwari, Syed S. Hassan, Fazal Hanan, David W Ussery, Antaripa Bhattacharya, Debmalya Barh, Artur Silva, Anderson Miyoshi, Vasco Azevedo. **Computational comparative genomic based insights into human gastric pathogen *Helicobacter pylori* (38 species), essential features and species pangenome**. (Manuscript. 2013)

My Contribution to the papers

All the work presented here was performed under the direct supervision of Prof. Dr. Vasco Azevedo, in the Laboratory of Cellular and Molecular Genetics (LGCM), Department of Genetics, Federal University of Minas Gerais, Belo Horizonte, Brazil.

- I. I participated in designing theme of the manuscript, sought literature for relevant data and performed the analysis. Co-authors (ARS, EB, SMB, SSH AM and DWU) helped me in restructuring the manuscript and their inputs in sequencing, alignment, assembly and annotation chapters. SCS and DB were involved in pathogenomic analysis and PPI, respectively. I wrote the first draft of the manuscript and corrected it after comments and suggestions from all the co-authors.
- II. I participated in the data collection and comparative genomic analysis. The technical guidance were obtained from DWU. SCS assist me in PIs predictions. All Co-author had put their inputs in the form of validation of results and discussion. I wrote the manuscript and finalized it after the suggestion and comments from all co-authors.
- III. I participated in planning, collection of data and management of the work and carried out comparative genomic analysis. To analyse this huge data access to CBS, DTU servers was provided by DWU. All co-authors contributed in writing up and approving the final manuscript.

There are other related articles in which I worked as the co-author. The list of these articles is given in *Curriculum Vitae*, appended at the end of the thesis.

I. PRESENTATION

I.I Research Groups, Collaborators and Funding

This work was conducted in the Laboratory of Cellular and Molecular Genetics (LGCM) at the Institute of Biological Sciences (ICB), Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil. Furthermore, it had the following active partnerships: the Laboratory of DNA Polymorphisms (LPDNA), Federal University of Pará; Center for Biotechnology (CeBiTec), University of Bielefeld, Bielefeld, Germany; Institute of Integrative Omics and Applied Biotechnology (IIOAB), India; Comparative Microbial Genomics Group, Center for Biological Sequence Analysis (CBS), The Technical University of Denmark (DTU), and Department of Mathematics and Computer Science (IMADA), University of Southern Denmark (SDU), Denmark. The researchers involved were:

- Prof. Dr. Artur Silva, researcher and group leader of the Laboratory of DNA Polymorphisms at the Institute of Biological Sciences, coordinator of the Network UFPA Paraense Genomics and Proteomics (RPGP), Belém, Pará;
- Prof. Dr. David Ussery, researcher and group leader of the Comparative Microbial Genomics group at Center for Biological Sequence Analysis (CBS) at Biocentrum-DTU, Denmark;
- PD. Dr. Andreas Tauch, researcher and group leader, Center for Biotechnology (CeBiTec), University of Bielefeld, Bielefeld, Germany;
- Prof. Dr. Jan Baumbach, researcher and group leader of Computational Biology Laboratory, Department of Mathematics and Computer Science (IMADA), University of Southern Denmark (SDU) Denmark;
- Dr. Debmalya Barh, researcher at the Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB) Nonakuri, India;

- Dr. Guilherme Corrêa de Oliveira, researcher and head of the Laboratory of Cellular and Molecular Parasitology, Research Center René Rachou - FIOCRUZ and coordinator of Minas Gerais Genome Network (RGMG).
- Dr. Stephen On, leader of Food Safety Program, Christchurch Science Centre, Institute of Environmental Science and Research (ERS), New Zealand.

I joined the Laboratory of Cellular and Molecular Genetics (LGCM), Department of Genetics in 2010 as full time PhD student under the supervision of Prof. Dr. Vasco Ariston de Carvalho Azevedo.

The research group (LGCM) has been actively involved with the characterization of the *C. pseudotuberculosis* genomes (bench and *Insilco*) for more than 15 years, even before the genomic era. Being the pioneer in *Corynebacterium* research in Brazil, the group engaged in intensive research projects covering diverse areas of biology, e.g. genomics, proteomics, transcriptomics. Also, the development of vaccines and diagnostics has made the group a reference point for the study of this microorganism.

The group also has gained sufficient expertise in the area of genomics with productive collaboration to the Laboratory of DNA Polymorphism (LPDNA), Federal University of Pará headed by the Prof. Dr. Artur Luiz da Costa Silva (Paraense Genomic and Proteomics Network – RGP). The groups acquired the high-tech sequencing facilities (SOLiD, Ion Torrent and Sanger: 3130 and 3730) and have successfully sequenced 15 *C. pseudotuberculosis* species so far, isolated from different locations around the world, biovars and hosts. The obtained sequenced genomes have been deposited to public databases (GenBank). Besides *Corynebacterium*, other species like *Campylobacter*, *Exiguobacterium* and *Leptospira* genome projects have been accomplished.

Since I did my BS and MPhil in Biotechnology and Molecular Genetics, my academic background was solely biology-based. Thus, I had limited computational skills and it was

initially challenging for me to work with bioinformatics group. However, my enthusiasm and devotion to the new area of science strengthened me enough to work with computational biology and bioinformatics, and to acquired sufficient knowledge in the field. During the initial period of my studies, I actively participated in ongoing sequencing projects, assembly and annotation of genomes. After learning and hands-on training in basic genomics, I got an opportunity to enhance my knowledge and acquired further skills in the area of comparative microbial genomics, while working at Center for Biological Sequence Analysis (CBS), Technical University of Denmark (DTU) under the supervision Prof. Dr. David Ussery, one of the pioneering and distinguished scientists in the area of comparative microbial genomics. This thesis was carried out with financial support from the following institutions: The Academy of Science for the Developing World (TWAS), Trieste, Italy (<http://twas.ictp.it/>), in agreement to National Council for Scientific and Technological Development (CNPq), Brazil in the form of a PhD fellowship. Sequencing facilities and assistance came from Foundation for Research Support of the State of Pará (Fapespa).

I.II Structure of the Manuscript

The thesis is divided into three parts based on three different bacterial species (*Corynebacterium*, *Campylobacter* and *Helicobacter*) selected for analyses. In the start, a brief history and introduction to genome and genomics is given, followed by strategies for comparative genomic, pangenomic and pathogenomic. A concise introduction to the selected genera (species) and their significances is given, followed by hypotheses and objectives of the study. The general methodology used in this study is demonstrated in a flowchart. The first chapter focuses on *Corynebacterium* beginning with introductory review on “Microbial Comparative Genomics”. In the first part, the review describes the developments and advances in genomic and comparative genomic sciences. In brief, an overview of microbial comparative genomics pre-requisites: sequencing technologies, alignment tools, annotation pipelines, databases and resources, visualization and genomic tools are described. Conclusively, the insights obtained from comparative genomic and pangenomic analysis and recent findings in genus *Corynebacterium* are discussed. The genus, as a whole, is considered as the model for analysis with particular interest in pathogenic species: *C. pseudotuberculosis*, *C. diphtheriae*, *C. ulcerans* and *C. kroppenstedtii*. Other published articles related to this chapter are mentioned in *curriculum vitae* at the end of thesis. Chapter 2 consists of a published article describing the comparative genomic/proteomic and pathogenomic analysis of the genus *Campylobacter*. Particular emphasis is given to pathogens *Campylobacter fetus* subspecies (*C. fetus* subspecies *venerealis* and *C. fetus* subspecies *fetus*), which are the causative agents of serious diseases like bovine genital campylobacteriosis (BGC), abortions, and gastroenteritis in human. Chapter 3 comprises a research article describing comparative genomic studies of genus *Helicobacter*, with specific focus on human pathogen *Helicobacter pylori*, the causative agent of peptic ulcer and gastric cancer in human population. The specific methodology, results and discussions are presented in the relevant articles (chapters). Additionally, a brief conclusion is given at the

end of each chapter. Also, a separate discussion chapter is provided where all the results are elaborated in the light of current literature. Furthermore, the potential features of the conducted study are highlighted.

The final conclusion and future perspectives of the work are presented in the last section. In the end, a *curriculum vitae* is affixed with academic and publications details. Additional (supplementary) files and figures associated with each chapter are provided in the appendices.

II. INTRODUCTION

II.I Genome and Genomics

The DNA (Deoxyribonucleic acid) was first isolated as early as 1869 and one of the remarkable discovery in science was the description of the DNA helix (nucleic acids) structure by James D. Watson and Francis H. C. Crick in 1953 (Watson & Crick, 1953). However, the term “genome” was first proposed by German botanist Hans Winkler, in 1920. He merged the letters “**gen**” from genes and the suffix “**ome**” from chromosome ((McKusick & Ruddle, 1987). According to modern molecular biology and genetics, the genome can be defined as “The entirety of an organism's hereditary information”. Genomes can be classified based on the organism or organelle which contain it, for example prokaryotic, eukaryotic, nuclear, mitochondrial and chloroplast genomes.

Genomics was established by Fred Sanger when he first sequenced the complete genomes of a virus and a mitochondrion. His group established techniques of sequencing, genome mapping, data storage, and bioinformatics analyses in the 1970-1980s (Kuska, 1998; The Sanger Centre, 1998).

The term “genomics” was later proposed by Dr. Thomas H. Roderick in 1986. He coined the term by suggesting it to a journal, which is now known as genomics (Kuska, 1998). As a result, the new discipline genomics was established. There are several definition of genomics often creating confusion and misunderstanding. The simplest definition can be described as “The Study of Genomes” (Kuska, 1998; Yadav, 2007).

In the late 1980s, after series of debates and discussions in the scientific community, the genome sequencing projects were officially started including human, mouse and other organisms (microbes). In the bacterial domain, the breakthrough was the first bacterial genome *Haemophilus influenzae* in 1995 (Fleischmann *et al.*, 1995), followed by *E. coli* genome in 1997. By the end of the last century, almost 39 bacterial species had been sequenced. Among the eukaryotes, *Saccharomyces cerevisiae* was the first genome sequenced in 1996 followed by the first draft of the human genome released in 2001 (The

Sanger Centre, 1998) and the final version of human genome was released in 2003.

Genomics studies are vast and may include the organization, function and evolution of genetic material at the level of the whole genome rather than individual genes. It analyses genetic material of living organisms. The systematic study of genome information can be applied to resolve questions in biology, medicine and industry.

Genome analyses can be studied both at transcription (RNA) level as well as at translation (protein) level. Similarly, the study of and analyses of transcripts is called transcriptomics and the study of proteins or proteome (proteins produces by an organism's genome) is called proteomics. In similar manner the suffix "*omics*" is used frequently in genomic literature referring to something big, demonstrating studies in life sciences dealing with large-scale data and information. Recently, the word "*omics*" is used with diverse biological processes (analyses) such as metabolomics, metabonomics, metallomics, lipidomics, interactomics, transcriptomics, spliceomics, neuromics, physiomics, predictomics, and so on (Yadav, 2007).

II.I.I Microbial Evolution and Diversity

Microbes (bacteria) provide an interesting opportunity to study the mechanisms of evolution because of their specific features: they harbor a previously unsuspected diversity even within species and populations, they are found either in small or very large population sizes, they can survive or grow in most unfriendly (extreme) environments from the inside of eukaryotic cells to hot springs or spaceships, and they frequently acquire and exchange genetic materials with distantly-related organisms (Abby & Daubin, 2007).

Like all other forms of life, bacteria also undergo evolution. However, the process of evolution is relatively slow and steady. The bacteria pass through selection processes and with time genetic changes occur in bacterial population. These changes may be results of mutations of existing genes during DNA replication or horizontal gene transfer. Bacteria can take up genetic elements either from their environment or from another bacteria through conjugation or transduction with viral elements (Barcellos, Menna, Da Silva Batista, & Hungria, 2007). On

the other side, genome reduction is also one outcome of bacterial pathogenic evolution. But not all bacterial pathogens follow this evolutionary trajectory because many maintain larger genomes and undergo frequent genetic changes, often accessorizing their core genome by incorporating new DNA. The changes in DNA can result in altering the function of the encoded protein molecules, which might be a new trait. If this newly acquired trait gives the organisms a selective advantage, it may be inherited or transferred again and eventually become dominant in the population.

If the DNA sequences undergo mutations, insertions or some other changes, these variations directly affect the coded protein. This is due to the fact that proteins molecules are more diverse in structure and functions, compared to DNA (Nishida, 2012). Another mechanism that plays a role in variation is the DNA duplication (Pallen & Wren, 2007). Among these phenomena, horizontal gene transfer is one of the most important factors in prokaryotes that causes variation in bacteria.

To estimate evolutionary relationships between different species and to avoid the labor and cost issues i.e., whole genome sequencing, single-gene analysis is widely in practice. However, due to some concerns about accuracy in phylogenetic analysis, phylogeny can be inferred from a number of universally conserved housekeeping genes using multi-locus sequence analysis (MLSA) (Urwin & Maiden, 2003). Although 16S rRNA gene sequence analysis and MLSA are recognized to be effective tools for phylogenetic analysis of bacterial species (Jolley *et al.*, 2012), a major limitation in these techniques is that only a small amount of information is used to represent an entire genome (Abby & Daubin, 2007; B. Trost, Haakensen, Pittet, Ziola, & Kusalik, 2010). These genes (16S rRNA) are sensitive to minor mutations and remain hot spots for variations (mutation) through evolutions ((Raskin, Seshadri, Pukatzki, & Mekalanos, 2006). These genes are also considered useful regulators to estimate the evolutionary relationships between organisms and to estimate the rate of evolution.

Beside these common principles of genome organization, comparative genomics has revealed a previously unexpected degree of diversity among prokaryotic genomes. One of the most striking examples of this diversity is the comparison of gene contents within and between species. In this manner, all forms of life seem to share only a handful of genes (~60), which are mainly dedicated to translation. The genes for other fundamental functions, such as DNA replication, transcription or basic metabolism seem to be more sporadically spread in the tree of life (Acencio & Lemke, 2009; Fraser *et al.*, 1995). Surprisingly, this diversity of genome content can be seen at every phylogenetic scale (Nishida, 2012).

II.I.II Characteristics of Microbial Genomes

In the last decade, development in sequencing technologies and bioinformatics tools has led to the characterization of many prokaryotic genomes in a very short period of time. The sequenced microbial genomes so far represent great phylogenetic diversity and many of them are associated with major human and animal pathogenesis. Microbial genomes compared to eukaryotic genomes show variations in structure and content density. For example viruses and organelles have tiny genomes sizes (kbs) and high gene density while bacterial genomes are small (Mb) with high gene density (Koonin & Wolf, 2008).

The sequenced bacterial genomes span two orders of magnitude in size, from ~180 kb in the intracellular symbiont *Carsonella rudii* to ~13 Mb in the soil bacterium *Sorangium cellulosum*. Although there are many genomes of intermediate size, the mentioned distribution suggests the existence of two (more or less) distinct classes of bacteria, namely those with 'small' and those with 'large' genomes (Koonin & Wolf, 2008; Raskin *et al.*, 2006).

The biological pressure and adaptation to diverse environments may also advance genomic difference for example, free-living bacteria possibly need to have more extensive adaptation potential, reflected by a larger genome, as they may come across more variable situations during their life, compared to pathogens (symbionts). Multiple DNA replicons (more than one

chromosome) can exist in bacteria. Additionally, some bacteria contain single or multiple copies of essential or non-essential plasmids (Nishida, 2012).

One of the genomic characteristics recommended for the description of species and genera is the G+C, where 5% and 10% are the common range found within a species or genera, respectively. This variation also modulates the amino acid content of proteins. In general, genome size is significantly associated with GC content, so that a higher GC content is more often observed for larger genomes and a low GC content is frequent in small genomes. Beside the physical nature, genome size and content, it is important to understand the functions of individual gene and protein, and characteristic differences among bacterial species (Nishida, 2012; Rottger, Ruckert, Taubert, & Baumbach, 2012). Despite the tremendous variety of life styles, as well as metabolic and genomic complexity, bacterial genomes show easily discernible and common architectural principles.

II.I.III Levels of Microbial Genomics

Genomics is a broad discipline, which may be divided into three main areas: **Structural genomics**, dealing with the physical nature of genomes. Its primary objective is to determine and analyse the genomic DNA sequence (mapping of genes – DNA sequencing – sequence annotation). **Functional genomics** is concerned with the way the genome functions. That is, it examines the transcripts produced by the genome and the collection of proteins and functional RNAs they produce (transcriptomics – proteomics). The third and relatively new area of genomics is **comparative genomics**, a large-scale, holistic approach that compares two or more genomes to discover the similarities and differences between the genomes and to study the biology of the individual genomes. This helps in identification of important, conserved portions of the genome and discriminate patterns in function and regulation. The data obtained also provide much information about microbial evolution, especially with respect to phenomena such as horizontal gene transfer (Ali et al., 2013; Jensen, Friis, & Ussery, 1999; Ussery, Wassenaar, & Borini, 2008).

In the last decade, introduction of next generation sequencing technologies and sophisticated bioinformatics tools have enhanced the genomic science and made it possible to sequence genome from diverse environments, faster and cheaper than ever. Consequently, the number of sequences on public databases is increasing at an exponential rate (Pagani *et al.*, 2012). This much genomic data greatly facilitated genomic science and established comparative genomics as a new discipline of genomics. In the next chapter, we will discuss in detail the next generation sequencing platforms, their specific features and associated bioinformatics tools and strategies for management of genomic data generated by these platforms .

II.I.IV Tree of Life and Statistics of Sequenced Genomes

Considering the tree of life and its three domains: Eukarya, Archaea and Bacteria and the statistics on Gold Online Database (www.genomesonline.org), the total number of complete genome projects is 4,169 (**Figure 1**). Among these projects, there are 183 and 182 complete projects from the first two domain (Eukarya and Archaea), respectively. On the other hand, from bacterial domain alone there are 3,804 complete projects and the sum total of genomes on the database are 23,423 (cited on march 6, 2013).

The number is rapidly growing and in the last couple of months we have witnessed a continuous increase in the sequencing projects. Phylogenetic distribution of sequenced genome recorded at end of last year (2012), among the bacterial genome projects (8,448) indicate that the highest number of species belong to Proteobacteria (46%) followed by Firmicutes (29%), and Actinobacteria (11%). All the three classes constitute 86% of the total bacterial genome sequencing projects (Figure 1). The remaining 14% contains bacteria from all other classes (Pagani *et al.*, 2012). Compared to eukarya and archaea, bacterial genomes projects species are sequenced at a much higher rate (Pagani *et al.*, 2012).

The interest in microbial genome sequencing is increasing because microbes provide an excellent starting point for studies as they have a relatively small and simple genomic

structure compared to higher, multicellular organisms. Furthermore, analysis of the microbial genomes also provides insight into microbial evolution and diversity beyond single protein or gene phylogenies. In practical terms, microbial genome analysis provides assistance to various biological phenomena, for example, environmental and waste clean up and wider applications in biotechnology. Most importantly, this analysis can be applied to designing new approaches for the treatment and control of pathogenic organisms (Forde & O'Toole, 2013; Metzker, 2005).

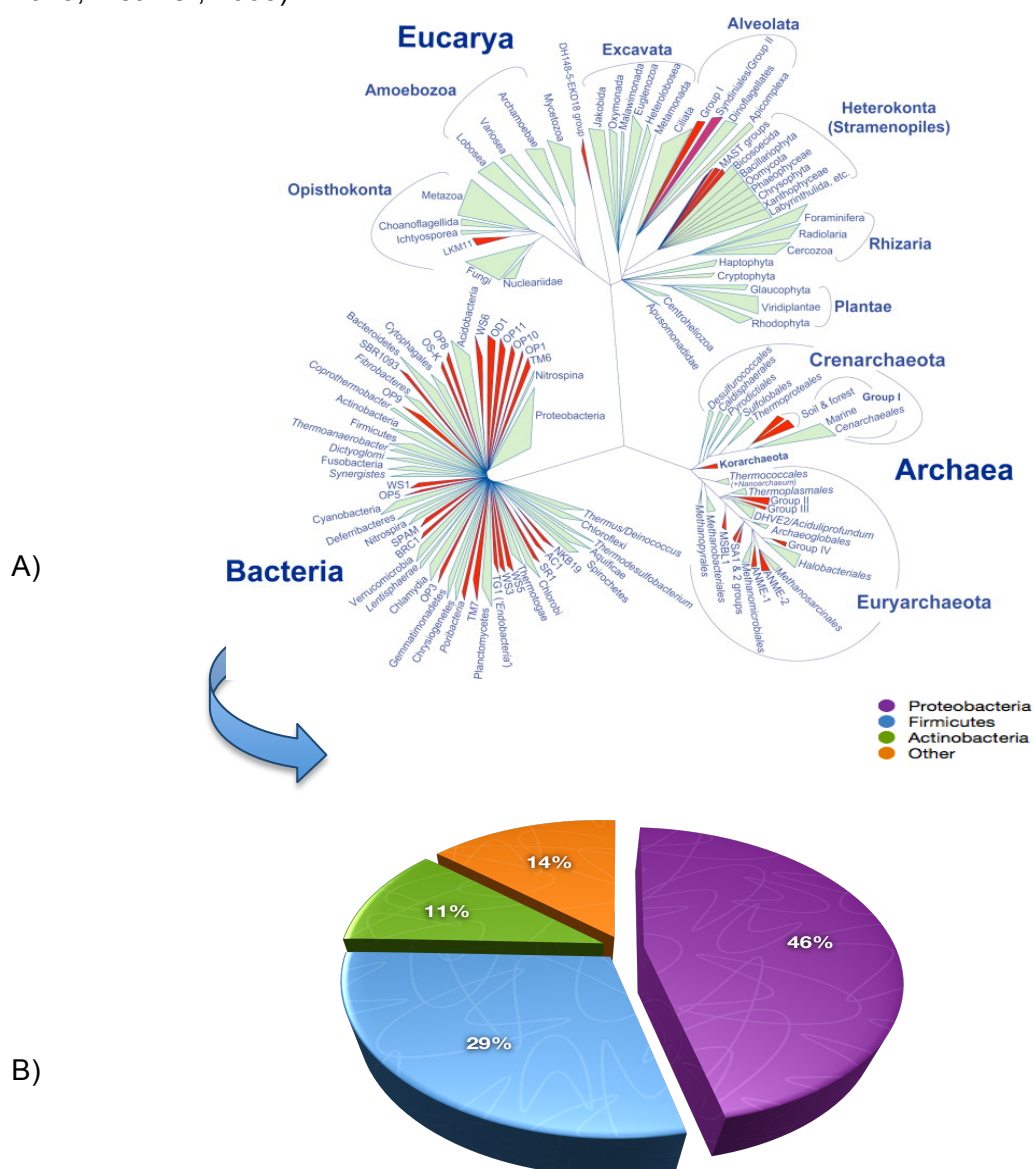


Figure 1. (A) Tree of life and its three domains Eukarya, Archaea and Bacteria.

(B) Phylogenetic distribution of sequenced organism.

(Data obtained from www.genomesonline.org - by March. 2013)

II.II Microbial Comparative Genomics

The availability of complete genomes is crucial to the entire enterprise of comparative genomics for at least two related but distinct and fundamental reasons: (i) the availability of complete genome sequences (complete sets of genes/proteins) provides the possibility to identify sets of orthologs, i.e. genes evolved from the same ancestral gene in the common ancestor of the compared genomes and also (ii) comparison of complete genomes (gene/proteins sets) is the necessary condition to determine not only which genes are present in any particular genome but also which ones are absent. The ability to define sets of orthologs and to pinpoint missing genes is crucial for genome-based reconstruction of an organism's metabolism, other functional systems and for reconstructions of genome evolution (Abby & Daubin, 2007; Grant, Arantes, & Stothard, 2012; Koonin & Wolf, 2008).

In post genomic era, multiple complete genome sequences from related and diverse species (pathogenic, non-pathogenic) have been published and are constantly added. Researchers are now interested in understanding the fundamental biological processes such as bacterial evolution, physiology and pathogenicity (Fraser et al., 2000). From evolutionary point of view, all modern genomes arise from common ancestors and information gained by one organism can have functions in others (similarly or differently). All the three levels of genomic analysis i.e. structural, functional and comparative genomic are equally important to decipher genomic sequences information. Differences between the strains of a distinct taxonomic clusters shows that the bacteria have a huge diversity. Therefore, comparison of multiple microbial genome sequences in the last decade has brought great insights into bacterial evolution and the amount of diversity (Raskin *et al.*, 2006).

The comparison of genomes with different life style will contribute significantly to the understanding of microbial evolution and will help deduce which genes are responsible for various cellular processes (Arnold & Jackson, 2011). The sequences from these organisms will also aid in our understanding of genetic regulation and genome organization. An

application of the genome sequencing is the complete genome sequence of *Mycoplasma genitalium*, a human parasitic bacterium and sexually transmitted organism with one of the smallest genome of 580 kilo bases in size and having approximately 517 genes. Among the interesting findings, minimum gene set required for laboratory growth conditions was estimated to be approximately 265 to 350 genes and about 100 of these have unknown functions (Mushegian & Koonin, 1996). This gave rise to the concept of minimal genome, i.e., the number (set) of genes required for an organism to survive. As an applications of comparative genomics, *M. genitalium* was compared to *Haemophilus influenzae* (larger genome having 1,743 genes). Interestingly, more than 40% of the genes were found to have unknown functions and also it was observed that the bacterium lacks three Krebs cycle genes and consequently a functional cycle. As soon the *Escherichia coli* K12 genome was published in 1997, when compared to others, it became evident that about 5 to 6% of the genes code for proteins are involved in cell and membrane structure, 12 to 14% in transport proteins, 10% in the enzymes of energy and central intermediary metabolic pathways, 4% in regulatory genes and 8% in replication, transcription, and translation proteins. Out of 4,288 predicted genes in this genome, almost 2,500 do not resemble known genes. These findings suggested deep insights of microbial genomic analysis and comparisons to understand microbial biology (Fraser *et al.*, 1995; Maurelli, 1998).

Different organisms share some common characteristics, and by comparison bacterial species can provides genomic information about the common essential biological process in different organisms. For example, identification of genes found in widely diverged species, can teach us what is required for fundamental biological processes of an organism (metabolism, biosynthesis etc.). On the other hand, the comparison of closely related species can reveal the origins of adaptive traits and events (virulence genes, pathogenicity islands, HGT etc.). In principle, genomes that diverged recently are expected to share more extensive gene order than distantly related ones. Comparative microbial genomics provides

the opportunity to understand fundamental information coded in genomic sequences by comparing them. It also helps in understanding the common (basic) and unique (special or common) genomic characteristics in closely related species. This would eventually lead us to the pathogens mechanism of evolution, host and environmental adaptation and more importantly pathogenicity (Pallen & Wren, 2007).

Comparative genomic analysis can be performed at different levels of the genomes to obtain multiple perspectives about the organisms. However, it can be divide into Intra-genomic and inter-genomic comparisons. The former analysis helps us to understand the degree of genes duplication, degree of divergence and genes organization in a genome. The later helps in estimation of degree of conservation between genes and degree of homology between genomes. Similarly, multiple intra- and inter-species similarities and differences can be utilized to classify the species based on genomic contents, rather than classical phenotypes and specific genomic regions can be identified for development of diagnostics and vaccine against pathogenic species (Kawai *et al.*, 2011; Röttger *et al.*, 2013).

II.II.I Gene/protein Analysis

To predict an evolutionary relationship and functional similarity between specific genes or proteins in different species/organisms, sequence analysis and comparisons tools are extensively used. Once the genomic sequences are available, *in silico* sequence comparison strategies are quite faster and cheaper. If a particular sequence (gene) is present in different organisms or conserved along evolution, it can be predicted that it might have a similar function in all the organisms. On the other side, based on observations, two molecules of related function usually have similar sequences reciprocally (Raskin *et al.*, 2006). However, searching diverse genes may show different evolutionary histories which reflects transfers of genetic material between species. By recognizing structure and function of a member of evolutionary family then it becomes possible to predict the function of all other member of that family. Computationally, it is convenient to compare two or more sequences for

discovering functional, structural, and evolutionary information in biological sequences (Wen, Wang, Li, Nie, & Yang, 2005).

Alignment of sequences from different origins is a tool to estimate the functional relationship between them. On the level of genes, the use of reciprocal BLAST hits is the most widely accepted approach suitable for tasks like gene annotation and the inference of homologies. If sufficient similarities in sequences is observed most probably they may have evolved from a common ancestor and the sequences are then defined to be homologous. Multiple sequences (more than one sequence) alignments often give us more information than pair-wise alignment because it is more informative about evolutionary conservation (Edgar & Sjölander, 2004).

II.II.II Intra-Genome Comparison (Single Genome Analysis)

Single genome sequencing is of practical importance in genomics and can bring enormous amount of information which leads to a broad range of knowledge about the organism (Fraser *et al.*, 2000). The analysis and comparison of individual genome sequences highlight events such as degree of genes duplication (multiple copies of a gene), degree of divergence and genes organization in a genome. For example, the use of alignment tools to find a specific gene in a genome is very fast and is employed to identify a genetic marker for a specific phenotype. The structural analysis of genomic DNA can locate chromosomal regions that lend themselves to certain genes and genomic elements. Beside this, bunches (clusters) of genes encoding surface-proteins (usually more AT rich), phage insertions, and regions expected to contain highly expressed genes can be identified (Friis, Jensen, & Ussery, 2000; Jensen *et al.*, 1999). Based on careful annotation of a genome it is also possible to find the gene neighbors of a specific gene, consequently identifying functionally connected genes. Therefore, to some major extent, the sequencing of individual genomes has facilitated the bench research. However, broader applications of genomics are manifested through comparative genomics. Even within a species, comparative genomics has highlighted a

diversity that would otherwise not have been detected (Fraser *et al.*, 2000).

II.II.III Inter-Genome Comparison (Pairwise Alignment)

Analysis and comparison of the global structure of genomes, such as nucleotide composition, orthologous genes, syntenic relationships, and gene ordering provides insight into the similarities and differences between genomes. Such comparisons provide information on the organization and evolution of the genomes, and highlight the unique features of individual genomes. The structure of different genomes can be compared at three levels: i) overall nucleotide statistics, ii) genome structure at DNA level, and iii) genome structure at gene level (Muzzi, Massignani, & Rappuoli, 2007).

To compare two genomes, exact functional counterparts of genes in genomes have to be identified and compared. Homologues are genes derived from some common ancestral gene. Paralogues (para = in parallel) are homologous genes comprising a multi-gene family (as a result of gene duplication) with possible variations in functionality. Due to the presence of many multi-gene families in genomes which are homologues, one has to identify the exact functional counterpart of a gene in another species out of a multi-gene family. These functional counterparts are called orthologues (ortho = exact) that have arisen from speciation. The orthologues are targets for comparisons and can be the only basis of gene comparison since the history of orthologous genes represents the history of species. To find the conserved genes in different organisms all-versus-all comparison of the genomes is performed using one of the three different strategies: i) a protein versus protein search; (ii) a DNA search of all the predicted ORFs of a strain against the genomic sequence of the other strain; and (iii) a translated protein search of all the predicted proteins of a strain against the complete DNA sequence of the other strain. A gene is considered conserved if at least one of these three methods produced an alignment with a minimum of 50% sequence conservation over 50% of the protein and/or gene length (He, Xiang, & Mobley, 2010; Tettelin *et al.*, 2005; Ali *et al.*, 2012).

There are many tools available for whole genome alignment and comparison for example, MUMmer and Artemis are available for comparative genomic analysis. These tools can be used for both pairwise genome alignment as well as multiple genome alignment. However, the BLAST (Basic Local Alignment Search Tool) algorithm, as well as other anchor-based algorithms, are commonly used for the identification of homologous gene candidates across diverse genomes (Lu et al., 2006; Tang et al., 2011).

II.III Pangenomics

Recently, when multiple complete genome sequences from closely related species have been obtained, it became evident that each new sequenced species brings in its novel/unique genes. Consequently, bacterial species have been observed with substantial diversity in them, resulting in the formation of new area of comparative genomics known as “pangenomics”, first introduced by Tettelin and colleagues in 2005 (Tettelin *et al.*, 2005).

The true diversity of a species can be estimated by multiple sequence comparisons across genomes calculating the pangenome of the species. Though hypothetical, these terms can serve to be used for defining and classifying bacteria. The so called pangenome contains the total number of genes found in the gene pool of compared genomes (Tettelin *et al.*, 2005).

The term pangenome can also be defined as “the global gene repertoire of a group of closely related organisms (preferably same species)”, includes core genome plus dispensable genome. The concept of pangenome can also be extended to other taxa such as genus.

Pangenome of bacterial species can be divided into different parts based on the conservation of gene content. The part that consists of conserved genes common to all genomes in a species (compared) is known as the core genome. It has been observed that core genome of phylogenetically coherent groups remains conserved in closely related species and contains genes that are less prone to Horizontal Gene Transfer (HGT). Hence, they are more stable, such as the house-keeping genes (Friis *et al.*, 2000). The genes

conserved between some of the species but not all species are categorized into dispensable genome. The third category called accessory genome, where the contents (genes) are confined to certain species (genomes), are also known as unique genes.

Pangenome can also be classified based on functions, and each gene in the pangenome can be classified into one of three groups (**Figure 2**): i) “core genes” which include those that control translation, replication and energy homeostasis (blue). The core genes often remain evolutionary conserved and have essential role in biological processes. Hence, they are good candidates for vaccine and drugs against pathogenic bacteria. ii) “character genes” are the set of genes (red) involved in colonization, survival or adaptation to a specific environment and are thought to form the lifestyle genes, which can also be named as the character genes (Röttger *et al.*, 2013). These genes define group or species and hence, are candidates for evolutionary studies. The third part is named as iii) “accessory genes” (genome) or the cloud as this part consists of genes that are rarely found in all of the species and are often strain specific, non essential and therefore less conserved (Lapierre & Gogarten, 2009). However, these genes can be used to distinguish strains or serotypes. These groups can be used to explain the differences and similarities between species or genera, and are visualized by pangenome trees which is also a part of this project (Snipen & Ussery, 2010).

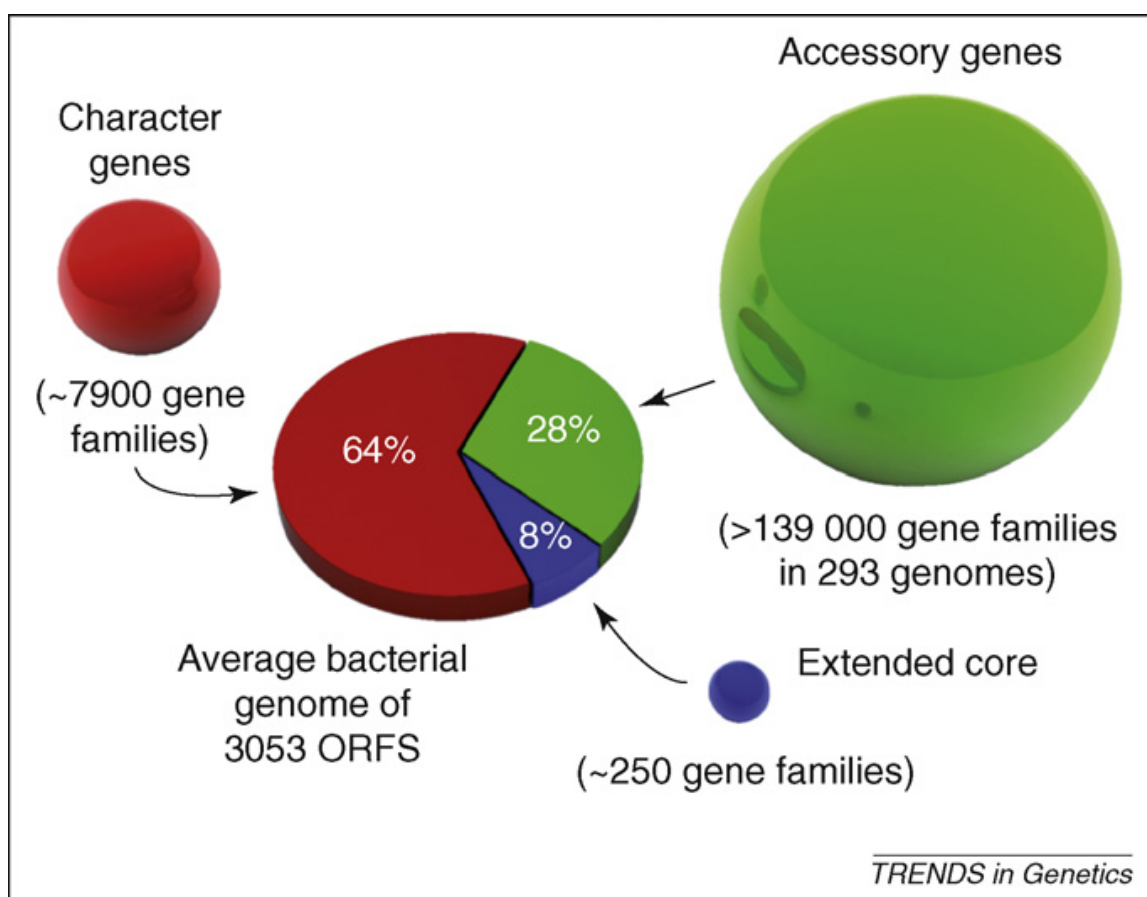


Figure 2. Illustration of the genes distribution in the pangenome of bacteria.

Results obtained by BLASTp analysis of 573 genomes and 15,000 sequences from 293 bacterial genomes. (Modified from Lapierre & Gogarten, 2009).

Bacterial strains from the same species may also vary considerably and pangenome of the species remains larger than the gene content of individual strains. Based on analysis of diverse and multiple genome sequences it has become clear that the bacterial domain has an open pangenome of infinite size. However, at genomic level, a typical bacterial genome may contains ~8% core genes, ~64% character genes and ~28% accessory genes (Lapierre & Gogarten, 2009).

II.III.I Open and Close Pangenome

Once the eight different isolates (genomes) of *S. agalactiae* (group B *Streptococcus*) were sequenced and compared for diversity estimation, the analysis resulted in addition of 33 new genes with each new genome (Tettelin *et al.*, 2005). The dispensable genes range from 200

to 300 (missing in at least one of the other genome). With emergence of new genes (strain specific genes) into the pool the species pangenome increases (open pangenome). Similar diversity was observed in case of *S. pyogenes* and 27 new genes were discovered with addition of each genome. Similar to the pangenome of Group B *Streptococcus* and *S. pyogenes*, the pangenome of *E. coli* (seven genomes) is open: the number of new genes contributed by each new *E. coli* genome is predicted to be 441 genes, substantially higher than the 27 and 33 genes predicted to be added with each new Group A and B *Streptococcus* genome (S. L. Chen et al., 2006).

On the other side, the close pangenome was observed while analyzing eight *B. anthracis* isolates. Upon addition of new genome, the novel genes added to pangenome continue to decrease and almost reach to zero (in case of adding the fourth genome). This is indicative of complete pangenome with only four genomes and hence poorly variable regions in genomes of *B. anthracis* (Muzzi et al., 2007).

II.III.II Application of Pangenome in Vaccine Design

The classical vaccine development approach requires the pathogen to be grown in laboratory conditions. The process of growing pathogen in laboratory is some times not that practical (highly infectious) and a number of limitations are associated with this approach i) the approach is time-consuming, ii) not applicable to non-cultivable pathogens, and iii) in many cases the antigens expressed in vivo during infections are not produced under laboratory conditions, or are variable in sequence (Muzzi et al., 2007).

The concept of reverse vaccinology was introduced by Rappuoli in 2000, where instead of doing the research for vaccine targets identification considering single strain or subspecies of an organism, simultaneously dozens of genomes could be explored for potential shared antigens and hence, pangenome reverse vaccinology was established (Lapierre & Gogarten, 2009; Rappuoli, 2000).

The determination of pangenome is of practical importance in identification of novel vaccine candidates. The core genes category represents the most appropriate sets for the selection of conserved and potentially universal vaccine candidates, furthermore, they are also more likely to be immunologically silent in any successful pathogen. On the other side, the set of dispensable genes, by contrast, might be an invaluable source of novel antigens, the reason is they are only present in a subgroup of selected genomes, although they might encode important virulence-associated functions and might be exploited in appropriate combinations to elicit a broad immune response. Beside the core genome, mobile genetic elements, such as transposons, plasmids and phage related genes, are frequently responsible for the pathogenic activity of bacteria or for drug resistance mechanisms. Generally, these genes do not belong to the core genome. Thus, their characterization will become more accurate as more genomes are sequenced.

The pangenomic approach was first applied on *S. agalactiae* (GBS) genomes by Maione and colleagues to design vaccine. The core and dispensable genome, were analysed computationally and putative surface-associated and secreted proteins coding genes were ascertained. Among the identified proteins, 396 were part of the core and 193 were part of the dispensable genome. Selected proteins (potential antigens) were then expressed as recombinant proteins, purified and tested for protection using an animal mouse model, in which mothers were immunized with the selected antigen, and the offspring are then challenged with the infecting strain. By doing so, four antigens were found capable of significantly increasing the survival rate among challenged infant mice. Unexpectedly, only one of these antigens belonged to core genome (not able to confer global protection, as its level of expression was shown to be highly variable among different strains). The remaining three were encoded by genes present in 75% of strains (dispensable). Nevertheless, the final vaccine was then formulated comprising a combination of the four antigens (universal strain coverage), levels of protection observed was similar to those when using capsular

carbohydrate-based vaccines (Tettelin *et al.*, 2005). This example demonstrates the importance of having access to the genome sequence of multiple strains and performing straightforward genome comparisons prior to developing vaccines based on genome data. Furthermore, as the number of genomes from related bacterial pathogens increases it would provide opportunity to select a suitable broad spectrum antigens for selected pathogens (Barh *et al.*, 2011a; Donati & Rappuoli, 2013).

II.III.III Essential Genes and Minimal Genome

In general, genes required for basic but fundamental cellular process (metabolism, biosynthesis etc.) are usually confined to the central (core) genome of the bacteria. The genes in bacteria which are indispensable for growth, cellular activities and foundation of life are known as essential genes. These genes constitutes a minimal gene (genome) set which is essential for an organism to survive (R. Zhang & Lin, 2009). The idea of essential genes came as a result of genomic analysis of a parasitic bacterium *Mycoplasma genitalium* (genome) containing only 468 proteins coding genes (minimal gene complement). When compared to *H. influenzae* 1703 protein coding genes, 240 *M. genitalium* genes were found orthologous to *H. influenzae*. Filtering the 240 genes the final set of 256 were suggested sufficient to survival of the *M. genitalium* (Mushegian & Koonin, 1996).

It has also been observed that essential genes remain conserved through evolution and are required by the organism for essential cellular processes such as metabolism and replication (Acencio & Lemke, 2009a). Few representative bacteria with estimated essential genes are shown in the Table 1.

Bacteria Species	Essential Genes
<i>Streptococcus pneumoniae</i>	244
<i>Mycoplasma genitalium</i>	381
<i>Haemophilus influenzae</i>	642
<i>Staphylococcus aureus</i>	653
<i>Vibrio cholerae</i>	779
<i>Escherichia coli</i>	1617

Table 1. Representative bacterial species with estimated essential genes

The identification of essential genes is of practical importance in drugs and vaccine development against bacterial infections. This is due to the fact that most antibiotics target essential cellular processes and essential gene products of microbial cells. Hence, these processes and products are candidate targets for such antibiotics (Acencio & Lemke, 2009). A database for essential genes, DEG (<http://tubic.tju.edu.cn/deg/>) is available for comparing the gene sets. DEG is widely used in *in silico* drug and vaccine development strategies and contains essential genes data from more than 10 bacteria, for example, *E. coli*, *B. subtilis*, *H. pylori*, *S. pneumoniae*, *M. genitalium* and *H. influenzae* (R. Zhang & Lin, 2009).

II.IV Pathogenomics

Pathogen evolution and host interaction is quite interesting. Generally, under stress conditions in the host, bacterial ability to uptake DNA is activated. This ability facilitates the pathogen to acquire genetic material that can help it resist the stress condition. Conversely, the loss of a mobile genetic element (MGE) carrying an avirulence gene can lead to enhanced virulence. The accessory genome, including MGEs such as genomic islands (GIs), plasmids, bacteriophage, insertion sequences and transposons, is often considered to provide the 'optional extra' genes (accessory genes, may be virulence factors) in a genome, enhancing the fitness and pathogenicity of pathogens (Arnold & Jackson, 2011; Jackson, Johnson, Clarke, & Arnold, 2011).

Comparative genomic analysis of pathogen genomes also provides new insights into how

pathogens have acquired common and divergent virulence strategies through evolution. By now, genomes of almost all major pathogens have been sequenced and analysis of sequences from diverse pathogens highlighted the crucial role of horizontal gene transfer and genome decay in the evolution of bacterial pathogens (Pallen & Wren, 2007).

Before the genomic era, only a few virulence genes were known in pathogenic species, leading to a merely incomplete sketch of the interaction of bacteria with the host. The determination of whole genome sequences of pathogenic bacteria and the comparison of genome sequences from pathogenic and closely related nonpathogenic species revealed a much more detailed picture of bacterial virulence. In some cases, it even invalidated previous concepts (Donati & Rappuoli, 2013; Santos *et al.*, 2011).

Comparative pathogenomics approaches have also enhanced our understanding of genomic modifications when comparing two or more genomes. Thus, they help us to build up the bacterial pathogen virulome (the virulence gene pool across species and genera) (Arnold & Jackson, 2011). Moreover, the predicted potential virulence factors by these comparisons and potential antigens identified by pathogenomic approaches will contribute in development of suitable vaccine. Determining the presence or absence of certain pathogenicity islands or genomic islands harboring multiple antibiotic resistance genes in bacterial isolates may further aid in identifying the cause of a disease, estimating its pathogenic potential, and predicting its antibiotic resistance. Thus, pathogenomic research has contributed to microbial diagnostics, pathotyping of bacteria, and the detection of novel target structures for the therapy and prevention of microbial infections (Hacker & Carniel, 2001; Schmidt & Hensel, 2004).

II.IV.I Genome Plasticity Analysis

Genome plasticity refers to the dynamic property of bacterial genome by which they involve in DNA gain, loss and rearrangement providing the microbe with higher adaptability to new environments and hosts (Maurelli, 1998). Genome plasticity can be generated by several

mechanisms like punctual mutations, gene conversions, rearrangements, inversion or translocation; deletions, and DNA insertions from other organisms, through plasmids, bacteriophages, transposons, insertion elements and genomic islands (Schmidt & Hensel, 2004). During such events i.e., recurrent losses of genes and functions must be compensated by the acquisition of new genetic material. In eukaryotes, the evolution of new genes is thought to occur mainly through duplication followed by sub- or neo-functionalization of one or both resulting copies. But prokaryotes can integrate genes of diverse origin into their genomes through HGT, which is believed to have a crucial role in speciation and prokaryotic adaptation to new environments (Koonin & Wolf, 2008).

Genomic Islands (GEIs) are large mobile elements which affect genome plasticity by carrying blocks of genes and causing evolution by leaps (Hacker & Carniel, 2001). GEIs may be classified, according to their gene content, into symbiotic islands, resistance islands, metabolic islands and pathogenicity islands (Barcellos *et al.*, 2007; Krizova & Nemec, 2010). There are several studies based on GEIs identification and their relationship with genome plasticity and also with pangenome size and singletons generation (D'Auria, Jiménez-Hernández, Peris-Bondia, Moya, & Latorre, 2010).

II.IV.II Pathogenicity Islands

Pathogenicity Islands (PAIs) are distinct genetic elements of pathogens encoding various virulence factors, acquired by bacteria through horizontal gene transfer (Dobrindt *et al.*, 2000). These islands are found both in gram positive and gram negative bacteria (pathogens) and are absent in same or closely related non-pathogenic bacteria. Characteristic features of these genomic regions include different pattern of GC content and codon usage from the rest of the genome and being often flanked by direct repeats (the sequence of bases at the two ends of the inserted sequence are the same). Presence of Insertion sequences or tRNA genes (sites for this integration event) acts as sites for recombination into the DNA. PIAs also carry some functional genes, such as integrases,

transposases, or part of insertion sequences to enable insertion into host DNA (Deng *et al.*, 2004; Friis *et al.*, 2000; Schmidt & Hensel, 2004). Well-known PIAs are in uropathogenic *E. coli*, the SPI-1 and SPI-2 islands in *Salmonella*, the Yop virulon in *Yersinia pestis* and the *Cag* island in *Helicobacter pylori* (Donati & Rappuoli, 2013).

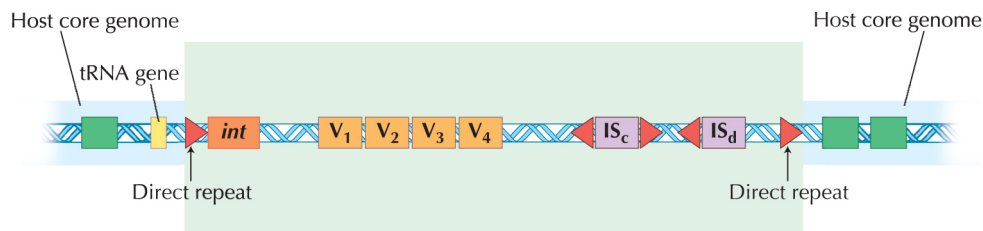


Figure 3. Graphical illustration of a pathogenicity island. The green shaded region (pathogenicity island) flanking by the host genes (blue shaded regions). On the left is a tRNA gene (in many of the cases islands are found near or within tRNA genes) at both sides of islands are direct repeats (red triangles). The orange boxes inside the islands are virulence genes and insertion sequences (purple boxes). Adapted from Schmidt H. *et al.*, *Clin. Microbiol Rev.* 17:14-56 @ 2004 ASM.

Some of the common virulence factors found on pathogenicity islands are adhesins, toxins, iron uptake systems, invasins, modulins, effectors and secretion systems (Type III and IV) (Zheng *et al.*, 2012). Identification of PAIs are important in understanding the causation of disease and the evolution of bacterial pathogenesis.

II.IV.III Virulence Factors in Pathogens

Virulence factors enhance the capability of the pathogen to cause disease in host. These virulence factors (VF) are characterized as potential targets for developing drugs. Inhibition of such virulence proteins would lead the pathogen avirulent as these proteins are the vital cause for establishment and severity of infection (Chakrabarty, 1998; L. Chen *et al.*, 2005; Deng *et al.*, 2004).

These virulence factors can be divided into seven groups based on their activities and outcomes in pathogenic lifestyle of bacteria: i) adhesins, it facilitates microbial attachment to host, ii) colonization, ability of the pathogen to colonize the host cells, e.g. production of enzyme urease by *H. pylori* to survive in the acidic environment of the human stomach iii) resistance factors (immune response inhibitors) or effectors that provide resistance to host defense mechanism, iv) invasions are factors which disrupt the membrane and facilitates process like endocytosis, v) toxins, factors produced by several bacteria, are harmful poisons to host cell and causes damages, for example, there are many food poisoning toxins produced by bacteria that can contaminate human foods and may cause food poisoning vi) polysaccharides (capsular) helps the pathogen in the protection from host responses for example, the polysaccharide capsule of *Streptococcus pneumoniae* inhibits phagocytosis of the bacterium by host immune cells vii) siderophores, for iron uptakes by pathogen (Zheng *et al.*, 2012). A number of virulence factor databases are available for public use including, **VFDB** comprising virulence factors of 24 genera of pathogenic bacteria, having option of comparison of virulence factors in related pathogenic bacteria on the database, and **MvirDB** a microbial database of protein toxins and virulence factors. (~64,711 proteins). The mentioned databases can be downloaded and can be used to align specific target sequences for confirmation (L. Chen *et al.*, 2005; Zhou *et al.*, 2007).

II.IV.IV Drug Targets Prediction

Identification of drug target against bacterial pathogen is considered as the first step in the drug discovery and development process. The availability of both pathogen and host genome sequences are the preliminary requirements. For any given pathogen, if both genomic sequences are present it becomes easier to identify drug targets at the genomic level. Recently, *in silico* target identification approaches have attracted researcher for its relative straightforward and efficient procedures. Hence, the strategies are gradually shifting from a generic approach to genomic and metabolomic approaches that are required to design new

defenses against antibiotic-resistant pathogens (Acencio & Lemke, 2009b; Barh, Jain, et al., 2011; Muzzi et al., 2007).

The human genome sequence greatly facilitated target identification research (Barh, Tiwari, et al., 2011; Haag, Velk, & Wu, 2011) and great advancement in the drug development has been observed from the time when human genome has been published. Use of computational approaches, with integrated genomics, proteomics, transcriptomics, interactomics, signalomics and metabolomics are current trends in target discovery for most human diseases such as cardiovascular, neuroendocrine, infectious diseases and more importantly for cancer. The reason is that these approaches make the discovery process faster and more cost effective (Barh, Jain, et al., 2011).

Currently, genomics and more specifically *in silico* comparative, subtractive and functional genomics are being extensively used to identify novel drug and vaccine targets in order to develop effective antibacterial therapies and vaccines against bacterial pathogens (resistant to antibiotics) or for which a suitable drugs and vaccines are not available (Barh, Tiwari, et al., 2011; Muzzi et al., 2007).

The general methodology for identification of drug targets includes: i) comparison of the pathogen sequences with host genomic sequences for homology and non-host pathogen sequences collection, ii) the analysis of these sequences for essentiality i.e., the target sequences should have role in essential biological processes in the cell and hence, be essential to pathogen. Furthermore, iii) mapping of sequences to pathogen and host metabolic pathways (common and unique to pathogen), and collection of proteins, enzyme and transporters etc. then iv) analysis of the targets for sub cellular localization in the cell, if it is found in cytoplasmic category and the product is an enzyme it may be utilized as suitable drug target. Conversely, v) if the proteins belong to membrane, surface exposed or secreted and are enzymes they may serves as both potential drugs and vaccines targets (Barh, Jain, et al., 2011; Barinov et al., 2009; Muzzi et al., 2007). However, for vaccine candidates

identification other strategies and sophisticated tools are available, which we will discuss in the following part (vaccine candidates).

II.IV.V Vaccine Candidates Identification

In the last few years, with the availability of multiple pathogen genome (related) sequences, *in silico* tools have become attractive to explore the whole genome sequences for universal vaccine candidates. As we discussed before, special attention is given to exported proteins because of their essential role in host pathogen interaction. By doing so, all the Open Reading Frames (ORF's) derived from the genome sequence can be evaluated with a computer program in order to determine their ability to be vaccine candidates. A number of parameter need to be considered for identification of true exoproteins (secreted and surface anchored), for example protein retention signals, lipoproteins, SEC pathway export motifs and transmembrane motifs. Vaccine candidates identification utilizes both pathogenic and related non-pathogenic bacterial genomes and for almost all major pathogens the genome sequences are now available on public databases (Barinov *et al.*, 2009; Donati & Rappuoli, 2013; Muzzi *et al.*, 2007).

To have a good vaccine candidate the predicted protein should have the following characteristics: i) evidence of being surface or secreted, ii) no multiple transmembrane helices (<2), iii) maximum adhesin probability (0.51), iv) ideal vaccine targets are those that exist in genomes of virulent pathogen strains but are absent in the avirulent strains and those capable of inducing strong immunity and avoiding autoimmunity, v) predicted vaccine targets are required not to have sequence similarity to proteins of host, vi) and MHC class I- and class II-binding epitopes. To optimize epitope vaccines, it has become an essential task to predict immune epitopes from protective antigens (Barinov *et al.*, 2009; He *et al.*, 2010). In contrast, cytoplasmic proteins are potential candidates for drug development. In chapters 2 and 3, the methodology has been discussed in detail while considering the veterinary pathogen *Campylobacter fetus* subspecies and human pathogen *H. pylori* genomic data

respectively.

II.IV.VI Reverse Vaccinology Approach

In the post genomic era, complete genome sequences from a range of pathogenic organisms became available. In recent past, comparative genomic and pangenomic analysis of multiple strains from same bacterial species revealed that genomic variability in bacteria is much more extensive than initially anticipated. It also became evident that even strains from single species show diversity and adaptability to different environments. Valuable information can be obtained from complete genome sequences and this has revolutionized the approach to vaccine development. The new approach starts with the complete information about the genome and the gene products and then identifies among these the important factors involved in virulence (Donati & Rappuoli, 2013; Rinaudo *et al.*, 2009).

The reverse vaccinology approach was first applied for the development of a vaccine against serogroup B *Neisseria meningitidis* (MenB), the major cause of sepsis and meningitis in children and young adults. The whole genome sequence is filtered by a number of bioinformatics tools to predict the potential Immunogenic proteins. Surface-exposed antigens are the most suitable candidates for vaccine development, due to their susceptibility to antibody recognition. Out of 91 novel surface-exposed antigens identified, 29 were able to induce complement-mediated bactericidal antibody response, a strong indication of proteins capable of inducing protective immunity (Rinaudo *et al.*, 2009). Later on, the reverse vaccinology approach was also successfully applied to other pathogens, including *Bacillus anthracis*, *Porphyromonas gingivalis*, *Chlamydia pneumoniae*, *Streptococcus pneumoniae*, *Helicobacter pylori* and *Mycobacterium tuberculosis* (Donati & Rappuoli, 2013; Rinaudo *et al.*, 2009).

II.V Introduction to Bacterial Genera and Species Selected

The following three genera: *Corynebacterium*, *Campylobacter* and *Helicobacter* are selected for comparative analysis and pangenome estimation. The particular interest in these genera is due to some pathogenic bacteria. Among the three genera, the genus *Corynebacterium* is selected as model for this study and data from various *Corynebacterial* species has been presented. In genus *Campylobacter* particular emphasis is given to veterinary pathogenic species *Campylobacter fetus* subspecies. Human pathogen *Helicobacter pylori* is prioritized from the genus *Helicobacter*. A brief introduction to selected species and their significance is given below and detailed information are provided in the relevant chapters of this manuscript.

The genus *Corynebacterium* is selected because of its interesting and important characteristics. To date, more than 80 species are taxonomically classified into the genus *Corynebacterium* and about half of them are associated with animal and human infections. Beside the pathogenic species the genus also represents biotechnological important bacteria and commensals of animals (Ott *et al.*, 2012). The *Corynebacterium* species are well studied and multiple complete genome sequences are available for comparative analysis. Members of the genus represent human pathogenic species (*C. diphtheria*), animal pathogens (*C. pseudotuberculosis*) and industrially significant organisms (*C. glutamicum*, and *C. efficiens*) (Röttger *et al.*, 2013). The interest in the genus is also because of species *Corynebacterium pseudotuberculosis*, one of our model organism. Recently our group has sequenced fifteen strains of *C. pseudotuberculosis* mainly from two biovar *ovis* and *equi* and the complete genome sequences are available on genbank. These strains were isolated from different hosts and different parts of the world (Brazil, Australia, UK, Egypt, India etc.) (Groman, Schiller, & Russell, 1984). We are therefore very much interested in comparative genomic based insights into the genus *Corynebacterium* and the genus pangenome (Ali *et al.*, 2013).

The *Campylobacter* species were selected for its substantial taxonomic diversity and wide host range (On, 2001). The clinical and economic significance of the genus is well

documented, since these bacteria are involved in a wide range of diseases affecting humans and animals. The Gram-negative, spiral-shaped bacterium, targets mainly cattle, swine, birds and is the main cause of human bacterial gastroenteritis (On, 2001; Skirrow, 1994). Among them *C. jejuni* is recognized as one of the main causes of bacterial foodborne disease in many developed countries. The *C. fetus* subspecies (*C. fetus* subsp. *venerealis* and *C. fetus* subsp. *fetus*) are prioritized due to associated serious diseases such as bovine genital campylobacteriosis (BGC), Infertility, abortions, septicemia and bacterial gastroenteritis. BGC is recognized as a significant threat to meat and dairy industries in Brazil, Argentina, Australia and New Zealand (Moolhuijzen *et al.*, 2009). Recently, our group and collaborators have sequenced a *Campylobacter fetus* subspecies *venerealis* strain. We, therefore, aim to compare our strain with that of *Campylobacter fetus* species *fetus* strain already on genbank to get insights into the species conserved genome and proteome and to identify regions associated with pathogenicity and candidate gene and proteins for diagnostic and vaccine development against the subspecies (Stynen *et al.*, 2011).

Finally, the global representative species (genomes) of human gastric pathogen *Helicobacter pylori* are selected to study in detail from the genus *Helicobacter*. Among all the species *H. pylori* is the widely known species of the genus for its worldwide prevalence. The organism remains a significant pathogen due to the fact that it is the most successful colonizer in the stomachs of almost half of the world's population (N. R. Salama *et al.*, 2007; N. Salama *et al.*, 2000). Particularly in human, *H. pylori* causes diseases such as gastritis and peptic ulcers, which may leads to the development of gastric cancer (~10% cancer deaths). It is therefore enlisted as a class I carcinogen by WHO, In 1994 (Dong, 2009; You *et al.*, 2012). The worldwide prevalence of *H. pylori* has been observed. However, the Infection (ulcers and stomach cancer) is more prevalent in developing countries compared to developed countries. Since its discovery in 1982, this species has been studied extensively but many features of the pathogenic lifestyle remain obscure and the clinical outcome of infection still cannot be

predicted from either bacterial or host genetic markers. We also observed that resistance to antibiotics is increased and limited attention is given to this important pathogen regarding development of drugs and vaccines (Dong, 2009).

Therefore, we attempted to collect genomic data from global representative species of *H. pylori* for comparative analysis and to understand the unique features and conserved regions in their genomes. We also aim to identify the candidate sequences for the development of diagnostic and treatment for *H. pylori* infections.

II.VI Hypothesis

Phylogenetically close bacteria share much of their genomic content and in case of pathogenic bacteria they are often confined to different hosts. By investigating their conserved core genome, it is suggested that common behavior between them can be characterized and we can have insights into mechanism of pathogenicity by identifying their unique genomic structural and functional characteristics.

To accommodate this, following hypotheses have been developed.

- Pathogenic species from the same genus are relatively less diverge from each other and from their non-pathogenic counterparts.
- Estimation of core genome for multiple related genomes will reveal the level of diversity and degree of conservation in their genomic contents.
- By comparing pathogenic genomes with that of non-pathogenic, it would result in estimation of genome plasticity (gain or loss) and pathogenic genes can be identified.
- Finally, the data can be used as input for literature search and databases for comparing and novel information can be exploited in the form of diagnostics, drugs and vaccines.

By implementing these hypotheses to this study for selected organism we aimed to find appropriate answers for the following fundamental biological questions and predictions:

i) what part of the genome remains conserved across and between bacterial species and genera; ii) what distinguishes pathogenic species from non-pathogenic species; iii) how related species share their genomic content; iv) what is the amount of minimal essential gene set in species; v) the number of new genes added to the gene repertoire of a species upon additions of new genomes; vi) how many virulence factors are there inside a genome and how they are involved in pathogenicity and finally; vii) prediction of potential therapeutics, drugs and vaccines targets.

III. OBJECTIVES

III.I General Objectives

General objectives of the study include: the use of *in silico* comparative genomic tools to highlight diversity in related bacterial species, estimation of species conserved genome, pangenome and pathogenomic analysis of selected pathogenic species from the genera *Corynebacterium*, *Campylobacter* and *Helicobacter*.

III.II Specific Objectives

- Phylogenetic analysis of selected species based on 16S rRNA genes and diversity analysis based on genomic variations;
- *In silico* proteome comparisons between and among selected species, whole genome alignments and estimated percentage of shared (similar) proteome;
- *In silico* identification of conserved genes among species and construction of conserved gene families for estimation of species core genome and species pangenome;
- Characterization of core genome into functional biological process categories based on COG functional categories and functional annotations;
- Using in-house pipeline PIPS for prediction of specific regions in pathogenic species, synteny analysis, rearrangement events and to characterize genomic plasticity between the genomes;
- *In silico* prediction of proteins in subcellular locations and exo-proteome analysis for identification of potential immunogenic proteins;
- comparative pathogenomics, prediction and characterization of common and unique virulence factors in pathogenic species, seeking to identify their involvement in pathogenicity;
- *In silico* identification of pathogens specific core essential genes in pathogenic species, using reverse vaccinology approach to predict broad spectrum drugs targets, and vaccines candidates.

IV. METHODOLOGY

IV.I General Methodology

Basic Local Alignment Search Tool, BLAST, is used in this project for the proteome comparisons and pan- and core genome determination in related species (genus). BLAST is an alignment tool to find local similarities between nucleotide or protein sequences. It can be applied in a number of ways, including strait forward DNA and protein sequence database search, gene identification, genomic repeats within a large sequence and multiple alignments search, etc.

The BLAST procedures have been performed locally on the server, this is due to the volume of data, expected high throughput running and planned post treatment of the BLAST outputs. in case of large number of genomes (*Helicobacter* 46 genomes) to compared, BLAST analyses were distributed to a number of processer (CPU) and hence process time is minimized. Additionally, the severs provides the opportunity to operate with special applications and settings, and thus gives a higher degree of flexibility.

The Pan- Core- genome estimation is a genome comparison method to form clusters of orthologous genes within a selection of organisms. In this way the genes are either to be classified as conserved or novel when genomes are compared pairwise. Conserved genes are grouped into gene families, then the number of gene families within a selection of genome is usually smaller than the cumulated number of genes. Genes that does not find a match is put into their own gene families as singletons. This context is done by pairwise and gradually BLAST processing of the genomes of interest against each other including against themselves. The criterion for finding the conserved genes in genomes is explained as bellow, “In order for two genes to be considered as conserved and thus to belong to the same gene family, the translated sequences have to be at least 50% identical within at least 50% of the length of the longest gene”.

This prerequisite is also referred to as the 50/50 rule and is used frequently in this project.

IV.II FLOWCHART

CHAPTER 1

CORYNEBACTERIUM

1.1 Review Article

Microbial Comparative Genomics: An Overview of Tools and Insights into the Genus *Corynebacterium*. *J Bacteriol Parasitol*. March 2013.

The manuscript highlights the recent development and advances in genomic science after the introduction of next generation sequencing technologies into the market. The strategies and tools in genomic research and importance of comparative microbial genomics have been discussed with emphases on genus *Corynebacterium*. In the start, basic genomic topics for example genome sequencing, alignment, assembly and annotation of genomes are discussed. As a model for study, diverse species from the genus *Corynebacterium* are selected for comparative analysis. Among the species, *Corynebacterium pseudotuberculosis* is an important means of dissemination of veterinary infectious diseases, *Corynebacterium diphtheriae* causative agent of diphtheria in human, *Corynebacterium ulcerans* commensal in domestic and wild animals, *Corynebacterium glutamicum* and *Corynebacterium efficiens* is used in industries for production of amino acids. Total of 11 representative species comprising of 19 complete genomes from the genus *Corynebacterium* are analysed for their similarities and differences. The pangenome of the genus is analysed using some of the mentioned tools and techniques discussed in the manuscript to get insights into the genus knowledge. We also demonstrated and discussed our relevant previous published results in this article.

Microbial Comparative Genomics: An Overview of Tools and Insights Into The Genus *Corynebacterium*

Amjad Ali^{1*}, Siomar C Soares^{1*}, Eudes Barbosa¹, Anderson R Santos¹, Debmalya Barh³, Syeda M. Bakhtiar¹, Syed S. Hassan¹, David W Ussery⁴, Artur Silva², Anderson Miyoshi¹, Vasco Azevedo^{1*}

¹Department of General Biology, Federal University of Minas Gerais, Belo Horizonte, 31907-270, Minas Gerais, Brazil

²Federal University of Pará, Belém, 66075-110, Pará, Brazil

³Centre for Genomics and Applied Gene Technology, Purba Medinipur, WB-721172, India

⁴Center for Biological sequence Analysis CBS, Technical University of DK-2800, Denmark

Abstract

Next generation sequencing (NGS) made it possible to provide whole genome sequences of pathogenic and commercially significant organisms in limited time, and with minimal cost. Computational comparative genomics is necessary, given that we sequence thousands of organisms every day, but our follow-up knowledge is still very limited. Nevertheless, genomic information from a single genome is insufficient to provide insights into the life style and extended view of the gene pool of a species. Multiple genomes could enrich our understanding of the relatedness of, and variations in organisms. Consequently, comparative genomic analysis remains powerful tools for identifying the orthologous genes in species, presence and absence of specific genes, evolutionary signals, and candidate regions associated with pathogenicity. Furthermore, pangenomic strategies, together with subtractive genomics, help in highlighting the inter- and intra-species relationships, conserved core and, pan-genome for characterizing virulence factors, drug targets and vaccine candidates. In this article, we present an overview of microbial comparative genomics pre-requisites: sequencing technologies, alignment tools, annotation pipelines, databases and resources, visualization and comparative genomic tools, and strategies. Finally, we present comparative genomic and functional analysis based insights and recent findings in genus *Corynebacterium*.

Keywords: Annotation; Alignment; Comparative genomics; *Corynebacterium*; NGS; Pangenomics; Protein-protein interaction; Regulatory mechanisms

Background

Genomics is one of the fastest evolving disciplines of science, where the breakthrough was the first whole genome sequencing of *Haemophilus influenzae* in 1995 [1]. The initial lag phase of genome sequencing was overcome by rapid advancement in sequencing technologies, assembling tools and efficient annotation pipeline. In recent years, we witnessed an exponential increase in the number of whole genome sequences in public databases and, to date, there are about 4,127 complete genome projects available for scientific explorations, including more than 3,700 bacterial genomes [2-4]. The constant demand to develop sophisticated sequencing technologies, capable of producing sequences with accurate genomic data in a faster and cheaper way, led to the development of the Next-generation Sequencing (NGS) technologies. Since the release of NGS platforms in 2005, these are responsible for a tidal wave of genomic information [5-7]. Nevertheless, the genomic sciences have a constant demand for *in silico* strategies, in order to change the sequences information into formats that are useful and easy to exploit by researchers. The following two key stages count greatly in genomics: i) Quality of the genomic data (assembly and accurate annotation); and ii) Management of genomic data (databases and analysis) [3]. As starting point, as soon as the data is retrieved from sequencing machines, the usual strategy is to assemble longer "Contigs" from individual sequencing "reads"; a number of interactive tools works to close gaps between contigs; and the genomic sequences (draft or finished) are then subjected to gene (ORF) predictions tools (Table 2), to identify the genes encrypted in the DNA sequence. Automatic annotation pipelines are used to predict the structural properties of the putative coding sequences (CDSs), and to deduce functions of the encoded protein and RNAs (tRNA and rRNA). Automatic annotation pipelines were developed to chase the promptly generated sequences, and for prediction of

their biological functions in the cell. However, manual curation with sufficient biological knowledge of the organism is an important step to avoid incorporation of misleading information in the public databases. Nevertheless, there are still potential reservations in manual annotation strategies (annotation section) [8,9]. Furthermore, the burst of genomic data generated by modern sequencing technologies in the recent past and the exponential growth of new sequences have made databases imperative tools for genomic research due to storage requirements and the constant need for *in silico* analyses of data. [3,10,11]. Therefore, a variety of electronic databases were developed with different data and storage forms that are publicly available on the web. The available genome-scale databases serve greatly in data organization and full time availability of the genomic data to researchers and professionals. Various important databases and resources, along with their data form, usage and applications are shown in table 2. Sequence alignment and comparative genomic tools are highly desirable for their potentials in identifying orthologous genes in species, specific genes, evolutionary signals, and candidate genes associated with organism's pathogenicity, adaptability, and economic significances [12-15]. The pairwise sequence-comparison methods employed in BLAST and FASTA have done great job in discovering the evolutionary relationships and

***Corresponding author:** Vasco Azevedo, Department of General Biology, Biological Sciences Institute, Federal University of Minas Gerais, Av Antonio Carlos 6627, 31270-901 Pampulha, Belo Horizonte, Minas Gerais, Brazil, Tel: + 00 5531 3409-2610; Fax: +00 5531 3409-2610; Email: vasco@icb.ufmg.br

Received February 12, 2013; **Accepted** February 25, 2013; **Published** March 06, 2013

Citation: Ali A, Soares SC, Barbosa E, Santos AR, Barh D, et al. (2013) Microbial Comparative Genomics: An Overview of Tools and Insights Into The Genus *Corynebacterium*. J Bacteriol Parasitol 4: 167. doi:10.4172/2155-9597.1000167

Copyright: © 2013 Ali A, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

functions of thousands of proteins from hundreds of different species, and even today there are tools to compare megabase-scale sequences [16,17]. Comparative genomic analyses are important not only for distantly related genomes, but also for closely related genomes, because of their applications in health and industry. Therefore, whole genome comparative analysis could have numerous advantages in narrowing down the valuable genomic information and identifying candidate regions in genomes [12,18,19]. For comparison strategies, there is no standard criterion for how many genomes (gene and protein sequences) shall be initially compared, i.e. one can start from two to an unlimited number of genomes. Moreover, the comparative studies may be performed on intra- or inter-species level, using bacteria with similar or different lifestyles (i.e. pathogenic/pathogenic, pathogenic/nonpathogenic and nonpathogenic/nonpathogenic organisms), depending on the study objective [18,20,21]. Taking into account the importance of the comparative genomic studies for understanding the inter- and intra-species genomic variations, conserved core- and species pan-genome, protein-protein interaction and regulatory mechanisms, virulence factors and candidate genes/proteins, and its application in designing vaccines, diagnostics and drug development against pathogenic bacteria. We selected several *Corynebacterium* species (pathogenic and non-pathogenic) from the class *Actinobacteria*, as model to get insights into the genus *Corynebacterium*. At first, the description of the important steps in functional genomics (strategies and demands) and comparative genomic analysis based results, followed by the *Corynebacterium* species relationships will be presented in a comparative manner, aiming to bring some light into the genus knowledge.

Next-Generation Sequencing Technologies

The “first generation” sequencing technologies were based on Sanger method, which uses termination of synthesis using 2',3'-dideoxynucleotides (ddNTPs) by DNA polymerases [22]. This technology has dominated the market for almost two decades, and was responsible for the release of the first complete bacterial genome in 1995 [1,23]. This state-of-the-art technology was achieved with the automated Sanger sequence by ABI Prism 3700 (Applied Biosystems), however, despite all its technical improvements, the need for development of better and faster methods remained [22,24]. The first NGS platform developed by 454 Life Sciences (www.454.com) was released in 2005 [24]. In the following years, other platforms were introduced into the market following the same general principle, which is to randomly sequence the DNA template from all the genome by breaking it into small fragments, and connecting them to specific adapters to be read during the DNA synthesis. The use of this methodology rendered the name Massive Parallel Sequencing

to these new technologies [23]. Although they follow the same basic principle, the existing NGS differ from each other concerning the unique combination of template preparation, sequencing and image, which are in turn responsible for the differences in the data produced by each platform [25]. The NGS technologies commercially available today include 454 GS20 Pyrosequencing-based (a method of DNA sequencing which determines the order of nucleotides in DNA) instrument (Roche Applied Science), Solexa 1G Analyzer (Illumina, Inc.), SOLiD instrument (Applied Biosystems), Ion Torrent (Life Technologies), and new SMRT (Pacific Biosciences). The basic features of each platform are shown in table 1. The length of the NGS read is smaller than the Sanger, which is the reason why these technologies are known as Short-Reads Sequencers. While Sanger generates reads between 1,000-1,200 bases, currently NGS offers between 50 and 500 continuous bases. Recently, a new platform that generates reads with greater length than Sanger was announced. The SMRT platform from Pacific Biosciences promises to generate reads with lengths greater than 3,000 base pairs, on average, within stances of over 10,000 base pairs, which would greatly facilitate mapping and assembly of the sequences (<http://www.pacificbiosciences.com>). High genomic coverage plays an essential role for a precise assembly of the genome in NGS technologies, since they generate short reads. That situation could appear as a problem when the genome present higher repetitive content, as the short reads can align in multiple locations of the genome [23,26]. After the NGS reads are generated, they are aligned against a reference genome or assembled *de novo*, which is an important step for NGS successful assembly process [27]. The *de novo* assembly presents more challenges when compared to the assembly through reference genome, as it is almost restricted to bacterial genomes due to the size of the genomes [28]. The greater benefits from the NGS technologies will only be possible once informatics science advances in maximizing the interpretation and utilization of short reads, including alignment and assembly [23,25]. Despite many challenges, NGS emerges as a dominant genomic technology due to its lower price, in comparison to Sanger methodology and its multiple applications. Most important, these new platforms provide genome scale sequencing for individual laboratories, which otherwise, would only be possible in large centers. Although there are greater advances in NGS technologies, they are still in their early stages, and the development of efficient pipelines of data analysis is crucial to transform NGS applications into routine research [26]. Technology is in constant evolving phase and has efficiently sequenced several genomes. Complete genomes of closely related organisms allowed large scale comparative and evolutionary studies, which otherwise were almost impossible just few years ago.

Sequence alignment

Technology	Approach	Read length	Bases/Run	Company and Web Addresses
Automated Sanger sequencer ABI3730xl	Synthesis in the presence of dye terminators	Up to 900 bp	96 kb	Applied Biosystems www.appliedbiosystems.com
454/Roche FLX system	Pyrosequencing on solid support	200-300 bp	80-120 Mb	Roche Applied Science www.roche-applied-science.com
Illumina/Solexa	Sequencing by synthesis with reversible terminators	30-40 bp	1 Gb	Illumina, Inc. www.illumina.com
ABI/SOLiD	Massively parallel sequencing by ligation	Up to 75 bp	1-3 Gb	Applied Biosystems www.appliedbiosystems.com
SMRT	Single molecule real-time sequencing	2,200 bp on average	120 Mb	Pacific Bio Sciences www.pacificbiosciences.com
Ion Torrent	Massively parallel semiconductor sequencing	100 bp on average	Up to 10 Gb	Life Technologies www.invitrogen.com

Table 1: Next-generation sequencing technologies; approach, read length, run and web addresses.

Once the genome sequences of closely related organisms are available, a desirable task in comparative genomics is to align two or more sequences. Alignment of sequences helps in various studies like gene and genome evolution, gene duplication events, signal for gene loss, repeat inversion or translocation events and rearrangement in genomes. Whole genome alignment is a useful strategy for detection of polymorphism, synteny analysis and sequence mapping, while multiple genome alignment could be used for identification of conserved sequences and sequence variations. Moreover, multiple alignments also support protein domain/structure and phylogenetic studies [29]. Local sequence alignment could be used for sequence homology searches, identification of DNA or protein sequence (annotation), and anchoring a whole genome alignment. In this context, the alignment software tools had a significant enhancement in last decade, being now able to solve the challenging tasks from a pair of prokaryotic organisms in a couple of minutes [30], to a pair of eukaryotic organisms in a couple of hours, running in a conventional desktop computer [31]. Nevertheless, there is a consensus about the urgent need for even better sequence alignment tools. The situation has been pointed out by recent publications on renewed ancient's alignment tools, or a combination of them emphasizing the "glocal" alignment strategy [31-38]. The reason behind this consensus is that genome alignment study is the most common and useful strategy for detection of plasticity events (i.e. horizontal gene transfer, polymorphism, recombination, insertions and deletions). However, this is not adequately addressed by alignment algorithms available today [38]. The common alignment tools for aligning pair of larger sequences include: MUMmer [17], AVID, and WABA [16], while for multiple sequence alignment, the tools available include: MAVID [37], MLAGAN [35], MGA [16], and MAUVE [37]. However, pairwise sequence comparisons BLAST [26], FASTA and MUMmer are common programs used for having their countless applications in finding evolutionary relationships and protein sequence functions [17].

Assembly and annotation

As discussed earlier, high-throughput sequencing technologies provides huge and fast growing amount of sequence information. Subsequently, the crucial stage is assembly (process to aligned short DNA/RNA sequences into longer ones) of genome starts, where the sequences are filtered according to the quality of the reads, and then overlapped into threads, based on either *ab initio* approach (matches in the pool of acquired sequences are considered), or on a reference assembly (the novel readings are aligned based on their similarities, with a previously assembled genome/phylogenetically closed), it is also referred to as mapping assembly [16,23]. The most important step in NGS data analysis is successful alignment or assembly of short reads to a reference genome. There are programs (MAQ, ELAND, SOAP, BLAST etc.) for alignment and mapping short reads, and to maintain the quality score [27]. On the other hand, *de novo* assembly is even more challenging due to the short read lengths and small bacterial genome size [27,28]. Due to the fact that shortness of read lengths causes huge problems in the subsequent genome assembly, phase and impeding closing of the entire genome sequence; however, recently hybrid *de novo* strategy (combining De Bruijn graph and Overlap-Layout-Consensus methods) is implemented to assemble entire genome of *Corynebacterium pseudotuberculosis* strain I19 from short reads, using a reference genome by anchoring, and remaining gaps are then closed using iterative anchoring of short reads by craning to gap. In comparison to classical genome sequence assembly with the same data as input showed that, with the availability of a reference genome hybrid *de novo* strategy is more effective as more genome sequences

could be preserved [39,40]. Besides, hybrid *de novo* strategy, table 2 shows common representative assembly tools. Nevertheless, properly furnished (assembled) genome containing highly accurate and integral sequences of an organism could greatly contribute to further data-mining, and can substantially contribute to the improvement of the annotation standard of newly sequenced genomes by genome comparisons [6,23]. In general, bacterial annotation is based on sequence homology and transferring information from already curated (reference), and/or closest genome(s) to the newly sequenced genome. Therefore, the quality of annotation greatly influences the comparative genomic studies. As mentioned before, automatic annotation pipelines help greatly in minimizing laborious job and time for annotation. There are several on-line services (IGS, IMG, JCVI, IGS, RAST, xBASE, BASys), which are simple in use, require little time investment, and also there are program/pipelines (AGMIAL, DIYA, Restauo-G, GenVar, SABIA, MAGPIE and GenDB), which could be downloaded and run locally, also useful where confidentiality or protection of data is required [41]. Various gene prediction tools and automatic annotation pipelines have been developed so far and are used for accelerating the annotation process (Table 2). These pipelines have significantly reduced the time and labor; however, it may have propagated errors sometimes; therefore, careful manual curation by biologists is required. Strategies like continuous literature search for experimental results and the use of GO terms could improve protein description and reduce syntactic errors [8,9]. Furthermore concerns with automatic pipelines must be addressed to avoid error propagation to new genomes, and more importantly to databases (e.g. UniProt, KEGG etc.). Based on observations, genomes from the same species often contain inconsistencies due to usage of different pipelines and strategies by independent research groups. These variations could have minor, but considerable annotation contradictions, for instance: taxonomic differences and misspelling during annotation, UniProt contain the word "syntase" instead of "synthase", 128 times; several identical genes have different names and more than one product, 'tnp' has 151 different product names, 'tnpA' has 97 and gene 'int' has a total of 12 different product names across 17 *Salmonella* RefSeq entries [8,9]. Furthermore the term "Hypothetical protein" appears much frequent, referring that the predicted genes is with no known homologs and experimental functional evidences, meaning that they may be real genes or mistakes of prediction tools. Thousands of entries in UniProt have been assigned the products "Hypothetical", "Hypothetical protein" or "Conserved hypothetical". It would surely be helpful if conserved features, motifs and scores of unknown function are added to them, since they may be recognized as true candidate/genes in nearby future. It is also important to note that, while naming the gene products, the annotator should avoid the words: "domain", "motifs", "homolog", "gene", "like", "similar" etc. Product names like "bacteriophage replication gene" should be replace to "bacteriophage replication protein". As observed, the reference genomes helps greatly in annotation, but do not always remain the best candidate for annotating the subsequent genomes, as it may be outdated. Refseq genome should be updated, when new strains and experimental data for the species become available [8,42]. An example of updating the Refseq is *Corynebacterium diphtheriae* NCTC 13129, where the re-annotation of the genome was responsible for an overall genome update of 57%. Briefly, 370 proteins, which were previously annotated as "Hypothetical protein", now have more descriptive functions with improved virulence characteristics and information about plasticity events [12]. An example of an open reading frame re-annotated and corrected for proper orientation based on BLASTp similarity is shown in figure 1.

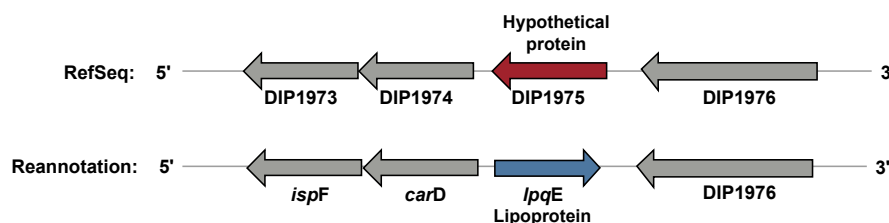


Figure 1: Re-annotation and correction of Open Reading Frame. ORF DIP1975 (red), in the wrong orientation annotated in the *C. diphtheriae* NCTC13129 Ref seq genome. The corrected ORF (DIP_1976) is illustrated (blue) with its probable genetic product, which was predicted based on searches for protein similarity (BLASTp), against the non-redundant protein database with cutoff: 10^{-6} . <http://dx.doi.org/10.2147/OAB.S25500>.

Graphical genome visualization and tools

Beside the universal genomic data storage distribution in XML format, the graphical and structural visualization of data is becoming common and useful mean for data exchange among researchers and the scientific community [43]. The genomic features represented in graphical maps provide structural characteristics of specific genomic regions on the chromosome, therefore, are easy to understand by the readers. Depending on the software and tool, structural features and number of functional features (annotation) can be obtained. To date, several open source and commercial software packages are available for creation and visualization of genome maps in linear, circular, or in both forms. In the last decade, Gibson and Smith [43] and Sato and Ehira [44] developed the programs “GenomePlot” and “GenoMap”, respectively, for generating genome maps (Atlas). Both are standalone programs, generating maps in different formats such as JPG, TIFF, GIF and PostScript. The GenoMap can also be used for map creation of other diverse data, such as microarray expression and gene localization data. However, interactivity, data input format and limited visualization options might be of major concern for some users, as the GenoMap is specifically designed for circular genomes [43,44]. To address the concerns in data visualization formats Kerkhoven et al. [45], present a web-based tool named Microbial Genome Viewer (MGV), for generating both linear and wheel maps with visualization of annotation and transcriptomic data. User can generate maps from provided annotation of uploaded custom annotations. For the visualization of complex data and high resolution images, the scalable vector graphics (SVG) format is used. Also, the Clusters of Orthologous Groups (COGs) functional categories, gene coloring option and data like GC%, GC- and AT-Skew can be visualized as colored gradients. Later in 2004, Stothard and Wishart [46] presented the CGView (Circular Genome Viewer), a Java application to generate both static and graphical maps, with zooming, feature labels and hyperlinking facilities. As the name indicates, CGView creates maps of circular DNA sequences, such as plasmids and bacterial genome. The information input can be done in three different types: Extensible Markup Language (XML); tab-delimited text files; and Protein table files, which typically end with “.ptt”, and are publicly available from NCBI ftp server. In all programs, PNG file format images are generated by default. However, JPG or SVG file formats may also be created through command line. However, the concerns remain about input files and viewer editable option. Genome Atlas Database, developed by Hallin and Ussery [47] in 2004, a web-based database, provides genome maps (Archaea and bacteria) with basic information like AT content, tRNA and rRNA counts, and more complex structural calculation. Another Interactive atlas, BacMap, developed by Stothard et al. [48], in 2004, uses CGView tool and generate high resolution, zoomable and color coded Images. BacMap also provides information regarding taxonomy, Gram’s staining,

chromosome numbers, physiology and relevance to host disease in tabular format. Later in 2008, Carver et al. [49], from Sanger Institute, UK proposed a Java application, “DNAPloter” tools, for creating both circular and linear genomic maps, with capacity of input file in common formats like GenBank, EMBL and GFF. All the presented software are robust tools in creating genome maps, however, they are offering comparative genomes visualization facilities. To address the issue, new tools, such as BRIG (BLAST Ring Image Generator) [50], Circos [51], and CGView Comparison Tool (CCT) [52], have been released recently. BRIG, an example of multiple genome comparison tools, is shown in genome plasticity and pathogenicity island prediction portion.

Genome statistics and dynamics

The genus *Corynebacterium* belongs to the class *Actinobacteria*, which are Gram positive bacteria with high G+C content. The genus contains about 80 species, which include commensal of human and animal, as well as pathogens (*Corynebacterium ulcerans*, *Corynebacterium diphtheriae*, *Corynebacterium pseudotuberculosis*, etc.) and industrially important bacteria (*Corynebacterium glutamicum*, *Corynebacterium efficiens*, *Corynebacterium variabile*, etc.) [53,54]. The life style of an organism is influenced by its basic genome statistics: number of chromosomes, numbers of coding regions (genes), gene density, GC and AT contents, and genomic signature (Oligonucleotide frequencies). Size of the genome (kbs, Mbs) varies among species, even among the strains of the same species. Biological pressures and environmental selection could also influence. Generally, the soil bacteria have bigger genomes compared to endo-symbiotic bacteria. It has been observed that many free living bacteria lose huge amount of their genomes, and while shifting from free-living organisms to symbiotic (pathogenic) [55-57]. Comparative genomics has revealed during comparisons between strains of related species, or/and species of bacterial pathogens, across the whole range of taxonomic variation, have made it clear that a ‘one size fits all’ approach cannot be applied to the evolutionary dynamics of bacterial virulence. Rather process like gene gain, gene loss and sequences change facilitates the variation. The smallest-scale variation, for example in bacteria (genomes), occurs at the level of single-nucleotide polymorphisms (SNPs). Its detection has been applied extensively to genetically uniform pathogens from the class of *Actinobacteria*, such as *Mycobacterium leprae* and *Mycobacterium tuberculosis* [58]. Nakabachi et al. [56] reported the smallest complete genome *Carsonella ruddii*, with circular chromosome of 159,662 bp, average GC% content 16.5%, an AT rich genome with high coding density (97%). Recently, Van Leuven and McCutcheon [57], the second smallest genome *Hodgkiniaci cadicola*, is reported with high GC content. There is consensus among scientist concerning the mutation rule that alters GC and AT proportions in genomes, and point mutation change the GC pair to AT much frequent than AT to GC [55,56]. Based

on observation, major change in GC content occurs in the third codon position; however, due to redundancy of genetic code, the nucleotide change in third codon position mostly does not alter the amino acid sequences. On the other hand, a significant increase in GC content of the first and second codon position results in changes in amino acid sequence of the encoded proteins. Besides, the highest AT content so far, observed in small genomes (insect nutritional endosymbionts) [14,59]. Consequently, the huge variation in bacterial GC content (13-75%) always attracted researcher and many assumed that the error in DNA replication biased is the key for the diversity. For example, the GC content ranges from 16% in *C. ruddii* to 75% in *Anaeromyxobacter*

dehalogenans, and these variations in GC content directly influences the genome size. It is also observed that GC content influences the codon usage, and for each 10% increase in GC content, the GC-rich codons increased by approximately 1% and amino acids encoded by AT-rich codons decreases by a similar scale [14]. For 11 species of *Actinobacteria*, the GC content is observed, which ranges from 42-74% (*Gardnerella vaginalis* and *Kineococcus radiotolerans*), and majority of the species goes around 60%, for phylum *Bacteroidetes/Chlorobi* ranges 22-66% and firmicutes found to be in range 23 to 68% [14]. However, a uniform GC percentage been observed in *Corynebacterium* intra-species, for example 6 species of *C. pseudotuberculosis*, which

Tool	Description/Features	Web Address/URL	Ref.
Assembly Tools			
CAP3	Alignment/assembly/Roche	http://pbil.univ-lyon1.fr/cap3.php	
Abyss	Alignment/assembly/Illumina	www.bcgsc.ca/platform/bioinfo/software/abyss	
Phrap	Alignment/assembly/Illumina/Roche	http://www.phrap.org/consed/consed.html	
Velvet	Alignment/assembly/Roche/ABI/Illumina	http://www.ebi.ac.uk/%7Ezerbino/velvet	
Gene Prediction Tools			
Glimmer	Microbial gene-finding system	www.cbcb.umd.edu/software/glimmer/	[102]
GeneMark	Gene Prediction in Bacteria, <i>Archaea</i> and Metagenomes	http://opal.biology.gatech.edu/GeneMark/	[18]
EasyGene	Gene Predictor in prokaryotic DNA	www.cbs.dtu.dk/services/EasyGene/	[43]
FgenesB	Bacterial Operon and Gene Prediction	http://linux1.softberry.com/	[103]
REGANOR	Gene prediction Server and Database	www.cebitec.uni-bielefeld.de/	[69]
Prodigal	Prokaryotic Dynamic Programming Gene finding Algorithm	http://prodigal.ornl.gov/	[104]
Automatic and Manual Annotation Pipelines/Tools			
GenColors	Comparative Genomics and Annotation Tool	http://gencolors.imb-jena.de/	[5]
MicroScope	Comparative Genomics and Annotation Platform	http://www.genoscope.cns.fr/	[6]
KAAS	KEGG Automatic Annotation Server	www.genome.jp/tools/kaas/	[23]
AutoFACT	Automated Annotation Tool	http://megasun.bch.umontreal.ca/	[25]
BASys	Bacterial Annotation System	http://basys.ca/basys/cgi/submit.pl	[42]
IGS	IGS Prokaryotic Annotation Pipeline	http://ae.igs.umaryland.edu/cgi/	[26]
CMR	Comprehensive Microbial Resource and annotation	http://cmr.jcvi.org/	[27]
PGAAP	NCBI Prokaryotic Automatic Annotation Pipeline	www.ncbi.nlm.nih.gov/genomes/	[28]
GenDB	Prokaryotic Genomes Annotation System	www.cebitec.uni-bielefeld.de/	[15]
MANATEE	Manual Functional Annotation Tool	http://manatee.sourceforge.net/	[41]
HAMAP	Automated and Manual Annotation of Microbial Proteomes	http://us.expasy.org/sprot/hamap/	[2]
RAST	Rapid Annotation using Subsystem Technology	www.nmpdr.org/FIG/wiki/view.cgi	[9]
xBASE	Bacterial Genome Annotation Service	http://www.xbase.ac.uk/annotation/	[42]
Blast2GO	Annotation and Sequence Analysis tool	http://www.blast2go.com/	
Databases and Resources			
NCBI	Genbank, RefSeq, TPA and PDB, databanks for storage and downloadable genomic information	http://www.ncbi.nlm.nih.gov/	
EMBL	Nucleotide Sequence Database	http://www.ebi.ac.uk/embl	
GOLD	Data resource for genomic and matagenomic projects	http://www.genomesonline.org/	
KEGG	An integrated database resource, provides genomic, chemical and systemic information	http://www.kegg.jp/	
IMG	Resource for Comparative Analysis and Annotation	http://img.jgi.doe.gov/	
JCVI	Comprehensive Microbial Resource (CMR)	http://www.jcvi.org/	
MBGD	Database, analysis of orthologous, paralogous, motifs, gene order and annotation.	http://mbgd.genome.ad.jp/	
RDP	Ribosomal Database, bacterial RNA sequences, alignments and tools for RNA analysis	http://rdp.cme.msu.edu/	
Rfam	RNA database	http://rfam.sanger.ac.uk/	
GtRNAdb	RNA Database, tRNA gene Predictions	http://lowelab.ucsc.edu/GtRNAdb/	
UniProt	Protein Resource and Functional information	http://www.uniprot.org	
UniProtKB	Curated Protein Database (UniProtKB/Swiss-Prot and UniProtKB/TrEMBL)	http://www.uniprot.org/help/uniprotkb	
Gene Ontology (GO)	GO Database, annotation of genes, protein and sequences.	http://www.geneontology.org/	
METACYC	Database for metabolic pathways	http://metacyc.org/	

Table 2: Gene prediction tools, an automatic and manual annotation pipelines, databases and resources. Tools for comparative genomics/proteomics analysis.

have 52.20% GC content in their genome in common, except the *C. pseudotuberculosis* CIP 5297 (52.10%) (Table 3). On the other hand, the AT content calculated for 11 species of genus *Corynebacterium*, including *C. diphtheriae* and *C. urealyticum* ranges from 32% in *C. variabile* and 47% in *C. pseudotuberculosis*. Moreover, intra-species genomes (*C. pseudotuberculosis*) been observed for negligible variation in their GC and AT contents. For example, *C. pseudotuberculosis* genomes remain stable for AT content (47% *C. pseudotuberculosis*). Interestingly, the genomes with similar GC contents found to have similar genomic signatures. Similarly, genomes with similar genomic signatures have similar GC contents. Nevertheless, Comparative genomics predicted that bacteria and *Archaea* have failed to gain horizontally transferred DNA with GC content higher than the GC content of their chromosomes. Therefore, the obtained DNA regions had lower GC content than that of the host chromosomal DNA [60].

Homologous proteins and whole genomes/proteomes pairwise alignment

In the post-genome era, determining groups of homologous proteins, (clusters paralogous and orthologous proteins), in bacterial species remains a challenge to bioinformatics. Protein sequences comparison is a powerful tool in characterizing the protein sequence for its preserved information through evolutionary process, and it is possible to identify proteins which share common ancestors, known as “homologous” [61]. The protein sequence comparisons are valued for identification of homologous proteins among species or genomes (and for many protein sequences evolutionary history could be traced back

to millions of years). As discussed before, with development of heuristic algorithms and powerful parallel computers, it is possible to have breakthroughs in sequence analysis based on homology. The routine and widely used program is BLAST (Basic Local Alignment Search Tool) [26], which allows the users to search for specific sequence(s) against the sequences in database, on the basis of homology with certain thresholds, and assigns each pair of proteins a similarity value. One step ahead, it is worthy to gather this data into groups (putative homologous proteins) by clustering tools, i.e. computational methods for partitioning data objects into groups, such that the objects share common traits, which have been measured with the similarity function. In the recent past, a number of tools had been developed for this purpose. Among them following tools have proven useful, and their accuracy is well studied: k-means, affinity propagation, Markov clustering and FORCE, as well as transitivity clustering (TC) [62]. The later strategy is applied to core genome of 89 actinobacteria, to find genes/proteins that are specific for certain actinobacterial lifestyles, i.e. different types of pathogenicity. With single intuitive density parameter, it is shown to be applicable for the task of protein sequence clustering.

Here, we selected and analysed a number of representative *Corynebacterium* species for homologies estimations, and literature data has also been sought for similar and supported results. The translated gene sequences in every *Corynebacterium* genome are compared by BLASTp (all-vs-all), against every other *Corynebacterium* protein in the dataset. The number of hits in a given set of proteomes is plotted against each other and the graphical matrix (blast matrix) for 11 selected *Corynebacterium* species is generated, which is shown in figure 2. The percentage identity between (any two genomes) genomes

Corynebacterial Species	Length bp	Predicted Proteins	%GC	% AT	tRNAs	16S rRNAs	Accession No.	Host/Source or Isolation	Disease/importance
<i>C. aurimucosum</i> ATCC_700975	2819226	2662	60.52	39.44	54	4	NC_012590.1	Human/vaginal swab/Germany	Pregnancy complication/ Abortion
<i>C. diphtheriae</i> NCTC_13129	2488635	2328	53.50	46.52	54	5	NC_002935.2	Human/UK	Diphtheria/1997
<i>C. efficiens</i> YS-314	3219505	2877	62.93	37.02	56	5	NC_004369.1	Soil and vegetable/Japan	L-glutamate and L-lysine producers
<i>C. glutamicum</i> ATCC_13032	3282708	3030	53.80	46.15	60	6	NC_003450.3	Soil bacterium/Japan	L-glutamic acid producer, 1950s
<i>C. glutamicum</i> R	3363299	3146	54.10	45.86	57	6	NC_009342.1	Soil/Japan	Industrially important
<i>C. jeikeium</i> K411	2476822	2137	61.36	38.64	50	3	NC_007164.1	Human/axilla/Germany	Nosocomial infections
<i>C. kroppenstedtii</i> DSM_44385	2446804	2128	57.50	42.54	46	3	NC_012704.1	Human/sputum/Uddevalla, Sweden	Patient with pulmonary disease/
<i>C. pseudotuberculosis</i> 1002	2335112	2138	52.20	47.80	48	4	CP001809	Goat/UFBA, BRAZIL	Abscess of CLA, 1971 ^a
<i>C. pseudotuberculosis</i> 42/02-A	2337606	2140	52.20	47.81	49	4	CP003062	Sheep/Dra Nicky Buller, Australia	Abscess of CLA
<i>C. pseudotuberculosis</i> C231	2328208	2139	52.20	47.81	48	4	CP001829	Sheep/ Dr. Robert Moore, Australia	Abscess of CLA, 1983
<i>C. pseudotuberculosis</i> CIP 52.97	2320595	2156	52.10	47.85	47	4	CP003061	Horse/Kenya	Lymphangitis, 1952
<i>C. pseudotuberculosis</i> FRC41	2337913	2139	52.20	47.81	49	4	NC_014329.1	Human/ Dr. Samer Kayal, France	Necrotizing lymphadenitis, 2006
<i>C. pseudotuberculosis</i> I19	2337730	2145	52.20	47.81	49	4	CP002251	Bovine/ Dr. Nahum Shpigel, Israel	Mastitis
<i>C. pseudotuberculosis</i> PAT10	2335323	2158	52.20	47.81	48	4	CP002924	Sheep/ Dra. Silvia Belchior, Patagonia	Abscess CLA, 2007
<i>C. resistens</i> DSM_45100	2601311	2272	57.10	42.90	51	3	NC_015673.1	Human/blood culture of leukemia patient	Multidrug resistant
<i>C. ulcerans</i> 809	2502095	2250	53.30	46.69	52	4	CP002790	Woman/Brazil	Pulmonary infection
<i>C. ulcerans</i> BR-AD22	2606374	2406	53.40	46.60	52	4	NC_015683.1	Nasal sample of dog/Brazil	Asymptomatic carrier dog
<i>C. urealyticum</i> DSM_7109	2369219	2011	64.20	35.81	51	3	NC_010545.1	Human/with alkaline-encrusted cystitis	Urinary tract infections
<i>C. variabile</i> DSM_44702	3433007	3175	76.10	32.85	59	6	NC_015859.1	Smear-ripened cheese	Uses in cheese industry

^a Caseous Lymphadenitis

Table 3: The *Corynebacterium* species selected for comparative genomic/proteomic and pathogenomic analysis.

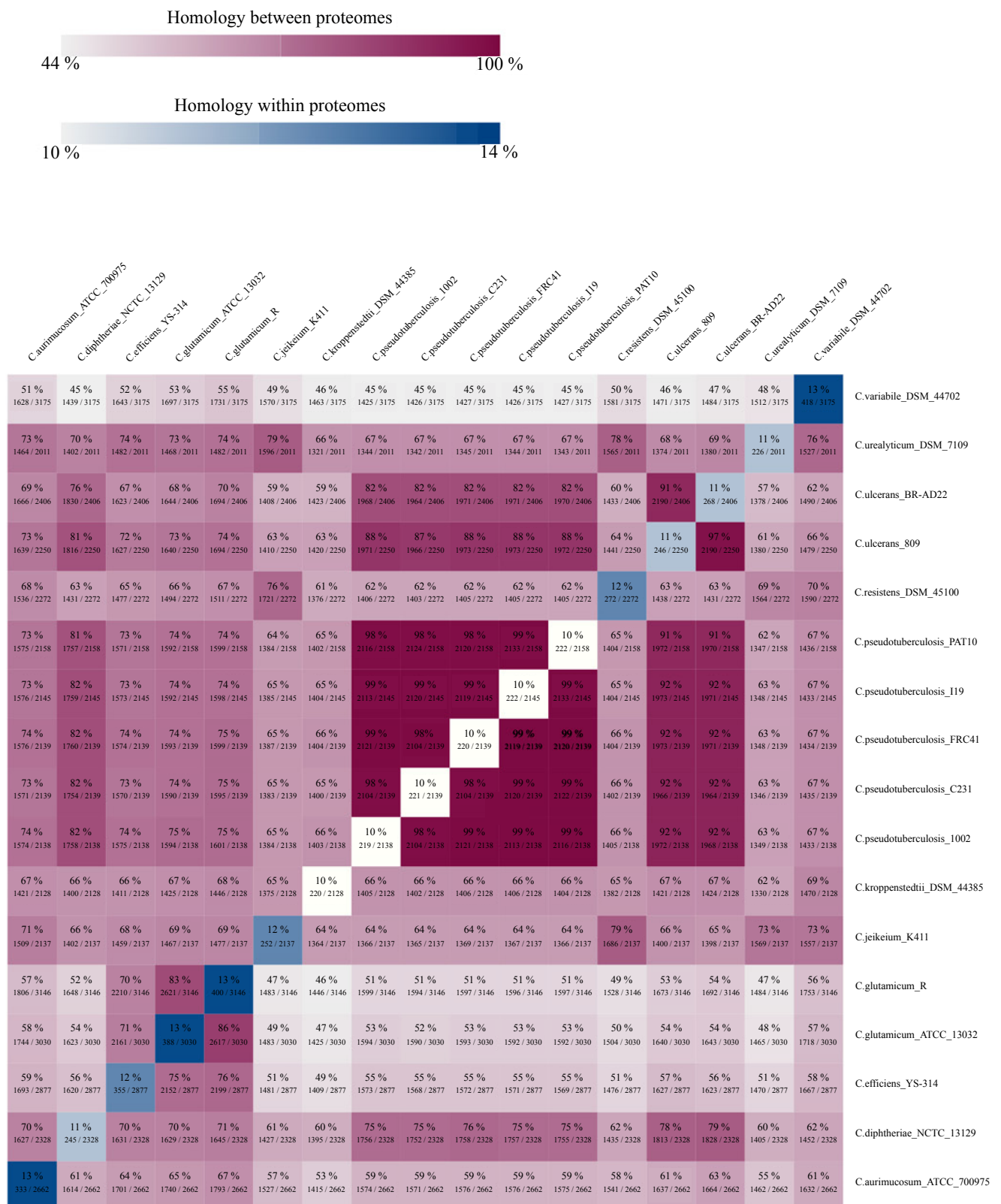


Figure 2: Pairwise genome comparisons. The Matrix illustrates number of conserved proteins and total number of proteins between any two species (pair wise). The color intensity is based on the relative percentage, darker the color, greater the conserved (homologous) proteins. Diagonal row in the matrix shows the internal homology within organisms' own proteome (percentages scale is given). Inter species highest homology can be observed in *C. ulcerans* and *C. pseudotuberculosis* (92%), while the intra-species conservation of gene families reaches to 99%, the highest in *C. pseudotuberculosis* (the dense square box in the middle).

combination by pairwise genome comparison is shown. The identity is expressed as the shared proteins (between any two genomes) divided by its total number of proteins, and visualized by color intensity in BLAST matrix (scale are given). Greater the intensity of the color indicates the highest fraction of genes/proteins found similar (homologous) between corresponding two genomes. As expected, a high BLAST score observed between intra- species (genomes), an indicative of a large fraction of shared proteins amongst them, for example, the highest similarity of 98-99% observed in *C. pseudotuberculosis* (Cp I19 Cp PAT10, Cp 1002 and Cp FRC41). On the other hand, the internal homologies in Cp genomes are up to 10% (homology within its own proteome). Among the *C. pseudotuberculosis* species, the lowest similarity (98%) observed between genomes Cp C231 and Cp FRC41, which is even higher compared to other species genomes isolated from sheep and human, respectively [63]. Despite of the fact that these species shows greater genomes/proteome similarities distributed to diverse hosts, mainly affecting small ruminant populations like sheep and goats, as well other mammals, for example bovines, pigs, deer, ovines, equines, rarely in camels and humans. However, they caused the same disease "Caseous Lymphadenitis" (CLA) or cheesy gland [63], which is highly prevalent in many regions of the world, resulting in huge and significant economic losses in agribusiness, since it is responsible for a decrease in wool production and carcass quality [64]. The species *C. pseudotuberculosis* also revealed greater inter-species homologies (92%), and remains closest to species like *C. ulcerans* (*C. ulcerans* BR-AD22 and *C. ulcerans* 809). The average similarities between the species been observed are 94%. As expected, the genomes of non-pathogenic specie, *C. glutamicum*, have the lowest similarity (46-75%) with other species of the genus. Intra-species proteome conservation (84%) is observed between *C. glutamicum* ATCC 13032 and *C. glutamicum* R. The species, *C. jeikeium* and *C. resistens*, were found to share 65% of their proteomic contents. Among the pathogenic *Corynebacteria*, the well known and most widely studied species is *Corynebacterium diphtheriae*, which is the causal agent of the disease "diphtheria" (upper respiratory tract illness). It shares 75% of their genome content with *C. pseudotuberculosis* species (causative agents of CLA), which is also considered to be taxonomically nearest organism (phylogenetic tree). Even so, it is a human pathogen and *C. pseudotuberculosis* is a veterinary pathogen, whereas in rare cases, causes disease in humans (*C. pseudotuberculosis* FRC41). The *Corynebacterium* comes in a so called group "CMN" (*Corynebacteria*, *Mycobacteria* and *Nocardia*), a group of pathogens having species with physiological and ecological heterogeneity, however, they share some common characteristics: a specific cell wall organization composed of peptidoglycan, arabinogalactan, mycolic acid polymers, and having high G+C contents in their genome [53]. From pathogenic point of view, among the *Corynebacteria*, *C. diphtheria* and *C. pseudotuberculosis* share greater conserved virulence factors. These factors facilitates the pathogen in various processes: Adherence, *srt* (A, B, C) and *spa* (C, D); Iron uptake, *fag* (A, B, C, D) and *hmu* (T, U, V), and *ciu* (A, B, C, D, E); and Regulation, *dtxR* (additional table 1). DtxR, the diphtheria toxin repressor of the human pathogen *C. diphtheriae*, is found conserved in all sequenced *Corynebacteria* until today. Over the last years, DtxR was subject to several genetic studies and the orthologous protein of *C. glutamicum* has been characterized [65]. *C. pseudotuberculosis* also share Phospholipase D (*pld*) gene with *C. ulcerans*, along with candidate virulence factors associated with process of adherence (*spa*) [21]. Details about an individual (any two species comparison) statistics, the number of proteins shared by any two species (strains), and the total number of protein, are mentioned in respective squares in the matrix.

Genome-wide Protein-Protein Interactions (PPIs)

The Sequence information is prior stage in understanding cells survival, reproducibility, behaviors and adaptation of organisms to various environments. One step further, the knowledge of about protein-protein interactions (PPIs) are vital in various biological processes, and are useful in determining functionality of uncharacterized proteins that are involved in critical events in bacterial survival, and/or pathogenesis [66,67]. Several ongoing researches tried to unveil genomic and proteomic information of various species, however, recently we reported *Corynebacteria* global protein-protein interactions [68]. For the first time, using a combination of comparative, functional, and phylogenomics approaches supported by published, experimentally validated data, we report (a) a probable conserved PPIs in the Cp proteome. (b) Further, we created proteome-wide common conserved PPIs for a number of pathogenic and non-pathogenic bacteria (*C. pseudotuberculosis*, *C. diphtheriae*, *C. ulcerans*, *M. tuberculosis*, *Y. pestis*, and *E. coli*). (c) Thereafter, the proteins involved in this common conserved intra-species bacterial PPIs were used to generate host-pathogen interactions considering human, goat, sheep, and horse as hosts. This host-pathogen PPI was based on experimentally validated published host-pathogen interactions data. (d) By analyzing the host-pathogen interaction networks, we identified common conserved targets in these pathogens. Analysis such as phylogenetic profiling [69], domain fusion [70] and gene neighborhood methods [71], have been used to develop genome wide PPIs in *C. glutamicum* ATCC 13032 and implemented to pathogenic species of *Corynebacteria*. The *C. glutamicum* ATCC 13032 genome having 2,993 proteins, generating a PPI of 5,476 interactions. A total of 1336 proteins are involved in these interactions, and 103 pathways can be mapped based on KEGG. In *C. diphtheriae* NCTC 13129 that has 2,272 proteins in its genome shows 5,293 interactions and 98 pathways. In *C. pseudotuberculosis* FRC41, which has 2,110 proteins in its genome, the number of interactions is 5,214 and pathways mapped are 97. However, common conserved genes/proteins of *C. pseudotuberculosis* FRC41, *C. pseudotuberculosis* 316, *C. pseudotuberculosis* 3/99-5, and *C. pseudotuberculosis* P54B96 when used, we obtained total of 4,186 interactions common to all these four *C. pseudotuberculosis* strains, and 68 pathways are mapped in this PPI. These four *C. pseudotuberculosis* strains, along with *C. glutamicum*, *C. diphtheria*, *C. jeikeium*, *C. efficiens*, *C. ulcerans*, and *C. glutamicum*, have 748 genes common to all. When we used these 748 proteins to make the PPI, a map having 2,794 interactions were generated, where 48 pathways can be found. Therefore it's obvious that the interaction varies depending on the species, and it's due the pan or core genome that is conserved phylogenetically [68].

Comparative functional genomics and systems biology (gene regulation)

Computational comparative functional genomics is necessary, given that we sequence thousands of organisms every day, but our follow-up knowledge is still very limited. Structural genomics helps in identification and descriptions of genomic DNA functional regions, however, information regarding regulation of these sites are of great importance in human medicine and molecular genetics [65]. Transcriptional factors (TFs) are DNA binding proteins, which influence or regulate the expression of target genes by binding to transcriptional binding sites, close-by the promoter regions. Some of the TFs may influence the regulation (up and down) of single gene, while others may do regulate various target genes. Nevertheless, cellular environment in or out, control the functionality of the these regulatory

factors [61,72]. Among the *Corynebacterium* species “*C. glutamicum*”, serves and model for the genus, however, for instance, <30% of the gene regulatory interactions are known. Considering the model *C. glutamicum* gene regulatory networks, an attempt is done to transfer gene regulations to human pathogens, *C. diphtheriae*, *C. jeikeium* and industrial relevant *C. efficiens*. By doing so, reliable transcription regulations are identified for about 40% of the common transcriptional factors, once there was very little knowledge about these regulations machineries [73]. For follow-up information regarding microbial gene regulatory interactions in *Corynebacteria*, ‘CoryneRegNet’ could be consulted, which is the reference database and as discussed above, beside *C. glutamicum*, *C. diphtheriae*, *C. efficiens*, *C. jeikeium*, and regulatory information are there. However, for other organisms, the databases and platforms could be helpful: RegulonDB, reference database for the prokaryotic model organism *E. coli*; MtbRegList, database for human pathogen *Mycobacterium tuberculosis*; PRODORIC, prokaryotic regulations database; DBTBS, database for Gram positive organism *B. subtilis* [65].

Comparative pangenomics (intra- and inter-species variations)

The term “pangenome” and its concept was proposed and described in literature for the first time in 2005 [74,75], where the term pangenome revealed the number of all essential genes present in a given group of organisms (the collection of all genetic material), preferably within the same species. Pangenome of a species could be further categorized into the core, dispensable, and unique genomes. The “core genome” (shared/conserved) usually contains essential genes for organism’s basic cellular functions, such as growth, reproduction, and survival. Moreover, the core genome is better representative of bacterial taxa at various taxonomic levels. The “dispensable genome” is the one, shared by few genomes in a set of genomes, where the genes are believed to have essential role in the genomic variation due to horizontal gene transfer, and the contents may have potentials for species-specific diagnostics, drug and vaccine development. The “Unique Genes” are those genes, which are confined to a particular strain (species). These genes may have involvement in bacterial critical activities of pathogenicity, drug resistance, and stress responses. Additionally, these factors may also increase the adaptability of pathogens to particular environmental conditions (free living bacteria), or hosts. However, they are not fundamental to the survival of the organism [62,75]. In principle, intra-species genomes must have larger conserved part, however, the gene content in species may differ considerably, and the pan-genome usually remains proportionally larger than the gene content of an individual genome. The core genome could be quite lower than the individual genome in the study. An example is the comparative analysis of four *Corynebacterium* species: *C. glutamicum*, *C. efficiens*, *C. diphtheriae* and *C. jeikeium*, it shows that all these species contain 1089 orthologous genes, which make up to 52% of all *C. jeikeium* K411 genes and 36% of the *C. glutamicum* ATCC 13032 gene complement [76]. Pangenomic studies are important in characterizing the species through the analyses of multiple strains genomes. However, the strategy of calculating the pan- and core genome could be applied to various sets of organisms, including intra and inter-species comparisons [75]. The study significantly extended to diverse organisms for their applications in genomic research; among them, *Bacillus cereus* [77], *Escherichia coli* [78], *Sulfolobus islandicus* [79], and many more examples can be found in recent literature.

In this paper, eleven species of *Corynebacterium* are analysed for

their pan- and core- genome estimations. The core genome is found to consist of 741 genes families and the pan-genome consists of 11,097 gene families. The observed pattern of new gene families into the pool is not uniform at the genus level. Where the core genome remains consistent (intra-species) or slightly decreases (inter-species) with addition of new species (genome), and the pan-genome is increasing substantially. The pan- and core- genome plots are generated and shown in the figure 3A. As described earlier, the core genome is significant part of a species and responsible for vital biological functions of the organism. According to Gene Ontology and its functional classification, at the third level of the biological process categories, the orthologous genes common to all species (core-genome) of the genus *Corynebacterium* have been classified and are shown in figure 3B. Based on our observation, if non-pathogenic species of *Corynebacterium* (*C. glutamicum* and *C. efficiens*), when kept a side the gene families, increases in the core- and consequently, the pan-genome size declines (data not shown here). On the other hand, the pathogenic *Corynebacterium* species (7 *C. pseudotuberculosis* genomes), with an average genome of 2,145 protein coding genes, shows uniform results, where the core genome consists of 1,660 conserved gene families (higher), and the pan genome consists of 2,296 gene families. An important finding which emerges from number of more genes into core genome of *C. pseudotuberculosis*, is the high similarity among the genomes. Since the results indicate a constancy of gene number, we expect, after the addition of more strains into the study, the core genome will be remain stable or might undergo a slight decrease. Based on this, no significant decrease will probably occur in the number genes in the core genome, and the number of genes families will remain constant. When comparing this data at genus level, a significant variation has been observed. Recently, we analysed intra-species pangenome of 15 *Corynebacterium pseudotuberculosis* species isolated from various host and geographical regions. Phylogenomic, pan-genomic, core genomic, and singleton analyses revealed close relationships among pathogenic *Corynebacteria*, the clonal-like behavior of *C. pseudotuberculosis* and slow increases in the sizes of pan-genomes. The resulting pangenome of *C. pseudotuberculosis* contained a total of 2,782 genes, which is 1.3-fold the average total number of genes in each of the 15 strains (2,078), and the core genome contains 1,504 genes, representing 54% of the entire pan-genome of the species (2,782 genes). Besides the species core genome (whole), the core genome of the *C. pseudotuberculosis* biovar *ovis* strains and *equi* contained 1,818 and 1,599 genes, respectively. The former shows more clonal-like behavior than later one, and most of the variable genes of the biovar *ovis* strains are acquired in a block through horizontal gene transfer, and are highly conserved [77]. Another example from the genus, genomic diversity and comparative genomic analysis of thirteen *C. diphtheriae* has shown to contain 1,632 conserved genes in the core genome and 4,786 in the pan-genome, with average increase of 65 genes per new strain addition in the studies. The number of core genes (70% of the gene repertoire) is considered higher than the non-pathogenic and pathogenic *Corynebacterium* species (*C. diphtheriae*, *C. jeikeium*, *C. efficiens*, and *C. glutamicum*), that showed conserved 835 genes. This phenomenon again supports the concept of same species isolates relatedness [80]. Generally, pathogenic strains from same species have little genomic variation in them, For example, two *C. ulcerans* (*C. ulcerans* 809 and *C. ulcerans* BR-AD22) strains, both genomes were found to be much similar, sharing (orthologous) 2,076 gene with a limited number of strain specific genes, which is due to a prophage-like elements in the *C. ulcerans* BR-AD22 chromosome. Also, there is a lower genetic rearrangement in the genus *C. ulcerans* 809. Furthermore, it is observed that, both *C. ulcerans* genomes are

more closely related to specie *C. pseudotuberculosis* (from 75-80% homology) than *C. diphtheriae* species (up to 50% homologous genes) [73]. Another comparative analysis of two pathogenic strains of species *Corynebacterium* (*C. pseudotuberculosis* 1002, isolated from goats; and *C. pseudotuberculosis* C231, isolated from sheep) showing greater similarity in their genomic architecture and gene content. Significantly, they revealed evidence of genome reduction, indicative of many genes lost, resulting in the smallest genomes in the genus. Features that could be part of the adaptation to pathogenicity include a lower GC content (52%) and reduced gene repertoire [62].

Genome plasticity and pathogenomics (virulence factors and targets)

Genome plasticity is defined as the dynamic property of bacterial genome that involves DNA gain, loss and rearrangement, rendering the microbe a higher adaptability to new environments and hosts [81]. Genome plasticity is generated by several mechanisms, like punctual mutations; gene conversions; rearrangements, as inversion or translocation; deletions; and DNA insertions from other organisms through plasmids, bacteriophages, transposons, insertion elements and genomic islands [82]. Genomic Islands (GEIs) are large mobile elements which affect genome plasticity by carrying blocks of genes and causing evolution by leaps [83]. GEIs may be classified according to their gene content, in symbiotic islands, resistance islands, metabolic islands and pathogenicity islands [84-86]. There are several studies based on GEIs identification and their relationship with genome plasticity, and, therefore with pangenome size and singletons generation [21,87-89]. Here, we have chosen *C. kroppenstedtii*, a pathogenic and lipophilic organism isolated from respiratory specimens of patients with mastitis [90], for illustrating that strategy. First, as *C. kroppenstedtii* is a pathogenic organism, we decided to search for pathogenicity islands (PAIs), a class of GEIs which presents a high concentration of virulence genes, appears associated to pathogenic bacteria, and is involved in the reemergence of several pathogens [91]. Second, to assess the variable genome content of *C. kroppenstedtii*, we have used a recently developed tool, called PIPS: Pathogenicity Island Prediction Software, which predicts PAIs based on specific features, like G+C and codon usage deviation; high concentrations of virulence factors and hypothetical proteins; and presence of transposase and tRNA flanking genes [92]. Third, we have chosen *C. glutamicum* NCTC 13032 as non-pathogenic organism of the same genus for genome comparison in PIPS. Finally, in order to generate a graphic visualization of the plasticity generated by PAIs, in relation to genomes of different species of the genus *Corynebacterium*, we have used the software BRIG: Blast Ring Image Generator [50]. PIPS have identified 17 putative PAIs on the genome sequence of *C. kroppenstedtii*. From figure 4, one can clearly see several deletion patterns on the other genomes, compared to the reference genome in the regions where the PAIs should be harbored. Those specific regions of *C. kroppenstedtii*, even though they can present high concentration of hypothetical genes, will account for the singletons of this species, and can be related to new functions and adaptability to new environments/hosts. Finally, in case they are advantageous for that species, they may be fixed on the core genome of the specific species, and/or transferred to other species of the genus, as exemplified by the presence of *Coryne* phage on the genomes of *C. diphtheriae* and *C. ulcerans*, and the PLD exotoxin coding gene (*pld* gene) in *C. pseudotuberculosis* and *C. ulcerans*, all of them harbored in PAI regions [62,93-95]. In similar comparative pathogenomic analysis, seven putative pathogenicity islands were predicted in two *C. pseudotuberculosis* (Cp 1002 and Cp C23), which contain signals of

horizontal transfer; the islands consists of several classical virulence factors, including genes for fimbrial subunits, adhesion factors, iron uptake and secreted toxins [62]. In addition to the above seven PAIs, when 15 *C. pseudotuberculosis* analyzed, a total of 16 pathogenicity islands (PAIs) are predicted. With respect to the gene content of the PAIs, the most interesting finding is the high similarity of the pilus genes in the biovar ovis strains, compared with the great variability of these genes in the biovar equi strains. Based on our findings, the polymerization of complete pilus structures in biovar *ovis* could be responsible for a remarkable ability of these strains to spread throughout host tissues and penetrate cells to live intracellularly, in contrast with the biovar equi, which rarely attacks visceral organs [77]. Among the pathogenic species, it is equally desirable to find the core and unique virulence factors in intra species. Therefore, proteome of two *C. ulcerans* (*C. ulcerans* 809 and *C. ulcerans* BR-AD22) species have been compared for pathogenic potentials and identification of virulence factors in them. Twelve candidate virulence factors (*rbp*, *cpx*, *pld*, *spa* (F,E,D,C,B), *rpj*, *cwlH*, *nanH*, *vsp1*, *vsp2* and *tspA*) have been identified with secretion signals and cell wall association [73]. Furthermore, a comparative genomic analysis of 13 *C. diphtheriae*, the diphtheria toxin gene “tox” was targeted in *C. diphtheriae* prophages and observed that *C. diphtheriae* Park-Williams No. 8 has been lysogenized by two copies of the *tox*⁺ phage and *C. diphtheriae* 31A carry unknown *tox*⁺ and DtxR (*tox* regulator detected by motif searches). Furthermore, the signals of horizontal gene transfer (subunits of adhesive pili) were also noticed in the pathogenicity islands predicted in *C. diphtheriae* [80]. We also attempted to find the targets for drugs development by subtractive genomic approach, in four *C. pseudotuberculosis* strains, Cp (CpFRC41, Cp1002, CpC231, and Cp119), along with CMN group of human pathogens. 20 conserved targets out of 724 genes (minimal genome) of Cp1002 are predicted. Two *Corynebacterium* specific (*mscL* and *resB*) and one broad-spectrum (*rpmB*) novel targets is proposed [53].

Overview of Ribosomal RNA and Pan-genomic Trees

The part of the DNA most commonly used for taxonomic classification of bacteria is the 16S rRNA gene, which could be compared among bacterial species and same for archeobacteria for variations. Evolutionary studies indicated that 16S rRNA genes continues to be sensitive to minor mutations, remain targets for variations, and considered useful evolutionary regulators to estimate the relationships between organisms, and the rate of evolution [96]. On the other side, the pan-genome is equally interesting in characterization of species or genus. It is also believed that low pan-genome diversity could be sign of stable environment, in contrast to a high pan-genome variation, which could reflect the considerable diversity in species and adaptation to diverse environments [97]. As an example, the trees based on 16S rRNA genes (extracted from the *Corynebacterium* species) and the pan-genomic family tree (based on the presence, and/or absence of conserved gene families among species) are compared, similarities in the distribution pattern of genomes in both trees are observed, the trees are shown in figures 5A and B. The observed pattern also supports the whole genome/proteome analysis shown in the blast matrix, where the closely related genomes from the same specie (*C. pseudotuberculosis* 98-99% homology), cluster together near to *C. ulcerans* specie (*C. ulcerans* 809 and *C. ulcerans* BR-AD22). According to the matrix results, greater similarity has been observed in the neighboring (taxonomically close) species (92%). Next to them, *C. diphtheriae* genome with 82% homology has an equal distance from both *C. pseudotuberculosis* species. Based on 16S rRNA genes sequences (homology) and pan-genomic analysis

A).

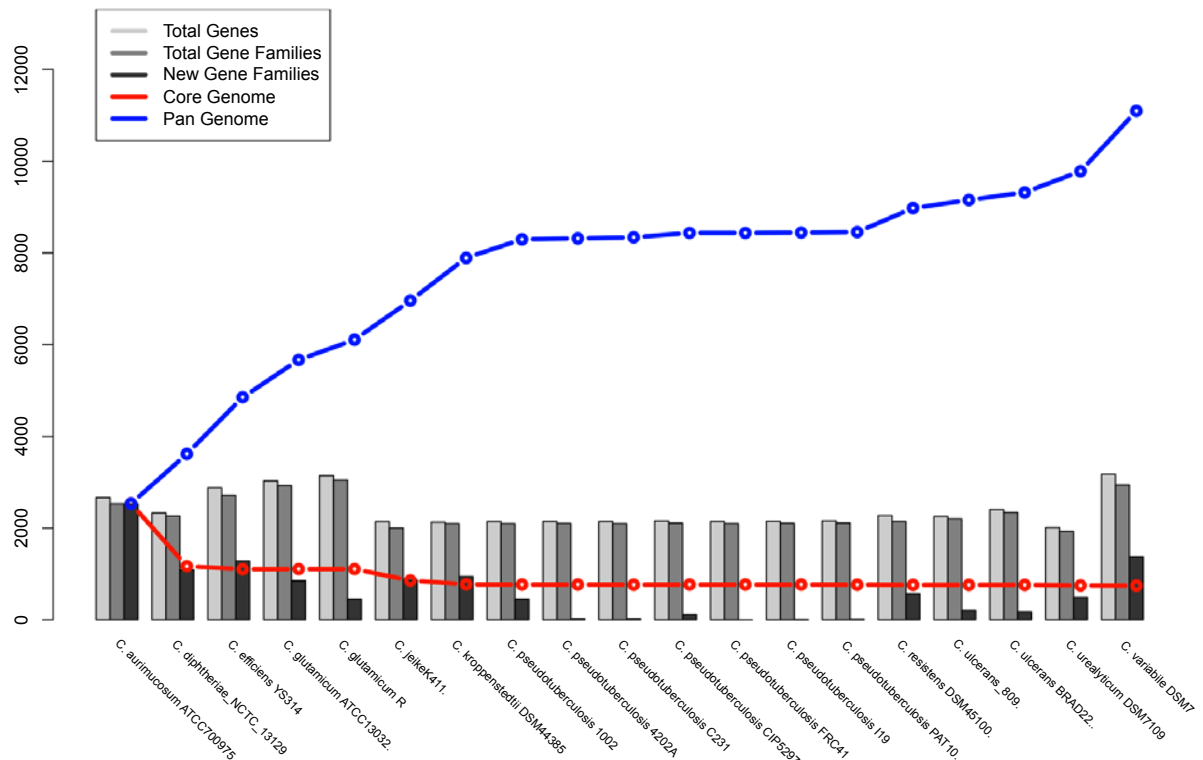


Figure 3A: Pan- and Core-genome analysis of 11 *Corynebacterium* species. The lines in blue and red represent the pan- and core genome, respectively. The pan-genome increases with addition of new species to the study (11,097 gene families), while the core genome decreases with slow rate, indicative of inter-species variations. From genome 8-12 (intra-species), core genomes remain almost stable, which demonstrate the greater similarity. For individual pair comparison and relatedness, BLAST matrix could be cited.

B). *Corynebacterium* Core Genome/Proteome Classification in Biological Process Categories

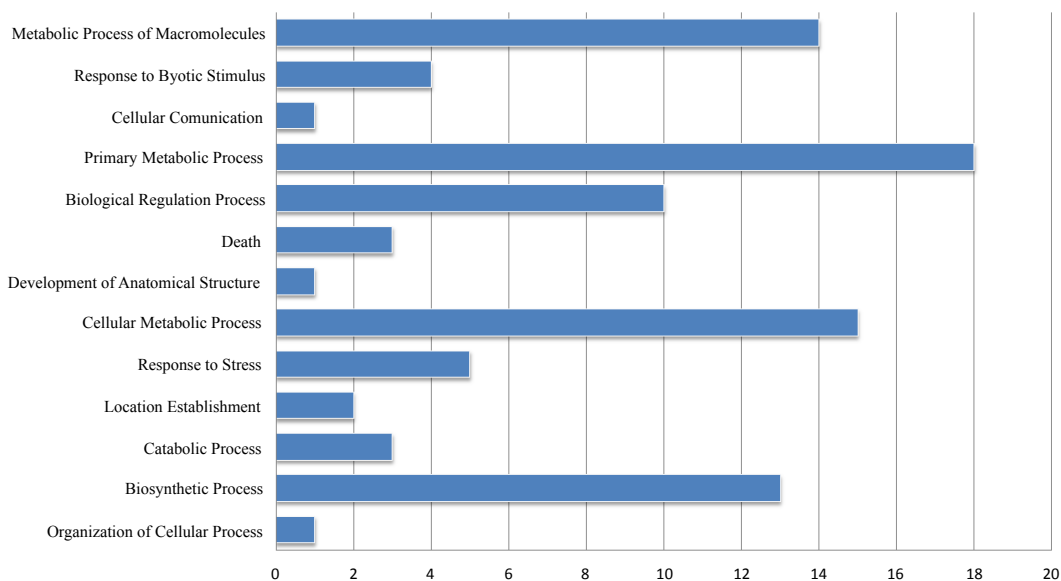


Figure 3B: The *Corynebacterium* species core genome classification in biological categories. The protein classification based on cellular function were performed by program Blast2GO (www.blast2go.org), majority of the protein found associated with metabolic and regulations processes in the cells. The conserved proteins may contain targets for broad range drug designing and diagnostics.

(conserved genes distributions), the *C. glutamicum* species (non-pathogenic and, of industrial importance) are cluster together and on separate clade with *C. efficiens* species. The overall picture of both trees and the data from blast matrix (proteome comparison) indicates that the results are comparable and the strategies could be used in parallel for evolutionary evidences and classification of organisms.

Multi-locus Sequence Typing MLST (and ribosomal MLST)

Multi-locus sequence typing (MLST) is an efficient tool for epidemiologic typing of bacterial pathogenic isolates. It was first developed by Urwin and Maiden [98], in 2003, and is based in the variation of core housekeeping genes observed after amplification and electrophoretic resolution. This technique can powerfully discriminate, and allows characterizing and classifying bacteria when appropriated number and function of genes are chosen. In order to type *C. diphtheriae* group, which includes *C. pseudotuberculosis*

and *C. ulcerans*, Bolt [99,100] developed a specific MLST. Isolates from different hosts of these Three species were type using primer combinations for assessment of inter- (genes in boldfont) and intra-speciesrelationship, as follow: *C. diphtheriae* (7 genes: *atpA*, *dnaE*, *dnaK*, *fusA*, *leuA*, *odhA*, *rpoB*);); *C. ulcerans* (6 genes: *atpA*, *dnaE*, *fusA*, *odhA*, *rpoB*, *pld*) and *C. pseudotuberculosis* (8 genes: *atpA*, *dnaE*, *fusA*, *odhA*, *rpoB*, *fagD*, *fagC*, *pld*). Species indicated no inter-relation, once no alleles were shared and evidence of recombination was not seen. MLST of *C. diphtheriae* strains was able to identify two distinct clusters formed by belfanti biotype and gravis, intermedius and mitis biotypes [100]. *C. ulcerans* strains from human and veterinary hosts showed to be genetically similar; and the biovars *ovis* and *equi* of *C. pseudotuberculosis* were genetically distinct, though are able to cause the same disease in different hosts [99]. MLST showed to be a useful comparative tool for typing *Corynebacteria*, and examine their relatedness and distinctness. Nevertheless, MLST analysis based on six to eight genetic loci not always give sufficient resolution among

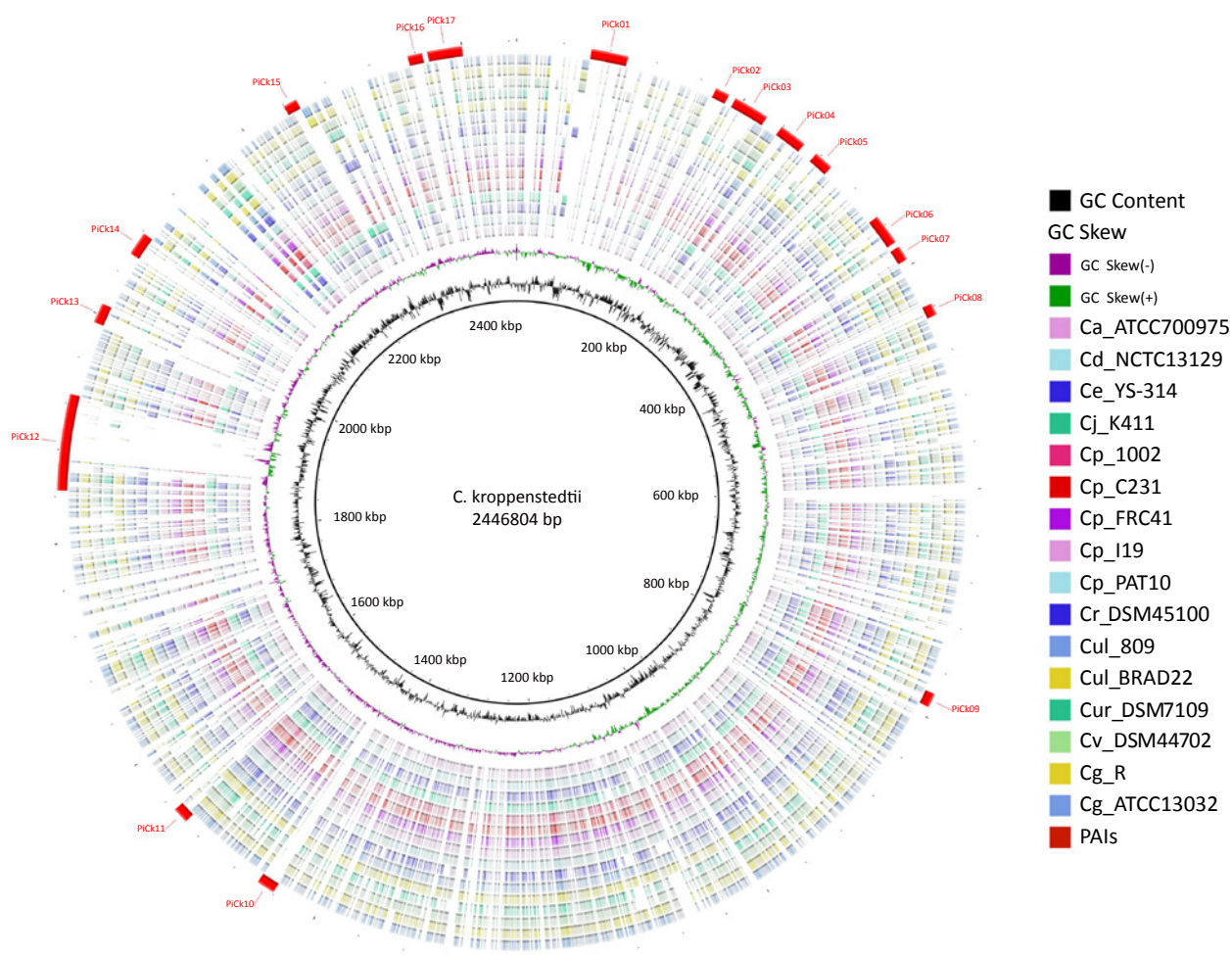


Figure 4: Genome plasticity in PAIs of *C. kroppenstedtii* compared to other *Corynebacterium* species. The figure shows the alignment of *C. aurimucosum* ATCC 700975 (Ca_ATCC700975); *C. diphtheriae* NCTC 13129 (Cd_NCTC13129); *C. efficiens* YS-314 (Ce_YS-314); *C. jeikeium* K411 (Cj_K411); *C. pseudotuberculosis* strains 1002 (Cp_1002), C231 (Cp_C231), FRC41 (Cp_FRC41), I19 (Cp_I19) and PAT10 (Cp_PAT10); *C. resistens* DSM45100 (Cr_DSM45100); *C. ulcerans* 89 (CuI_89) and BR-AD 22 (Cu_BRAD22); *C. urealyticum* DSM7109 (Cur_DSM7109); *C. variable* DSM44702 (Cv_DSM44702); and, *C. glutamicum* R (Cg_R) and ATCC 13032 (Cg_ATCC13032), using the genome of *C. kroppenstedtii* DSM 44385 as a reference sequence. The outermost circle highlights the seventeen putative pathogenicity islands of *C. kroppenstedtii* (PiCk 1–17) in red.

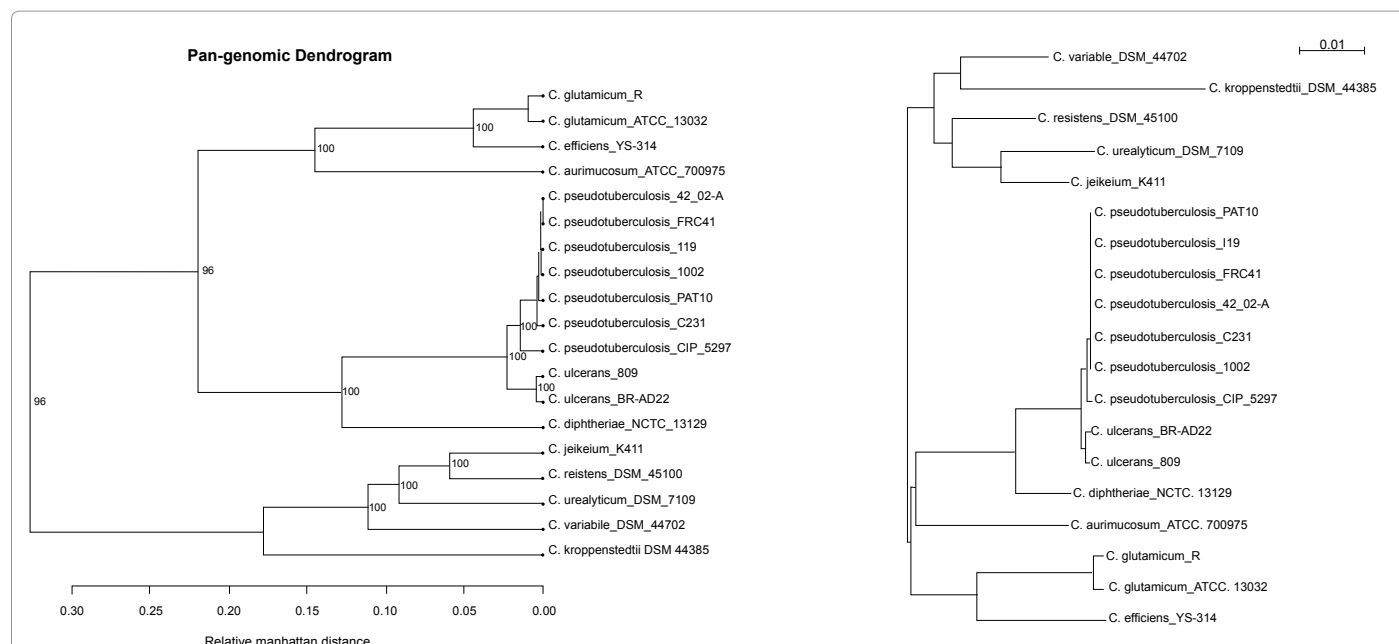


Figure 5A and B: The Pan-genomic tree and 16S rRNA based tree comparison. The Pan-genomic tree (left) constructed, based on the presence and absence of conserved gene families in species (genomes) and 16S rRNA gene sequence homology based tree (right). In comparison, both the tree shares considerable similar pattern of genome distributions to clusters. According to Blast Matrix and pan-genomic calculations, the *C. ulcerans* and *C. diphtheriae*, both pathogenic species are lies close to *C. pseudotuberculosis* (pathogens). The *C. glutamicum* non-pathogenic species clustered apart from pathogenic species based on both analyses. Hence, both strategies could be used in parallel for comparative analysis, typing and classification of bacterial species.

closely related bacteria, and each MLST scheme has to be developed for specific group of closely related bacteria. Recently, with the increase in available bacterial genomes, the demands for comparative analysis of the genetic variation in the shared loci become an imperative strategy. An alternative approach is Ribosomal Multilocus Sequence Typing (rMLST), an efficient computational analysis proposed by Jolley et al. [101]. The strategy is to target a larger set of genes encoding bacterial ribosomal subunits (*rps* genes) for microbial sequence typing. The significance of selecting the 53 ribosomal genes and *rps* loci for universal characterization includes its presence in all bacteria, distribution across chromosome and functional conservations. Based on *rps* loci variation, any bacterial sequence could be positioned from top at domain to bottom at strain level. The database (Bacterial Isolate Genome Sequence Database –BIGSDB) has developed, including 1900 complete bacterial genome and 28 draft genomes.

Conclusion and Future Perspectives

Genomics starts with sequencing, and sequencing techniques are evolving from Sanger's to NGS. Now, the limitations of short reads and the dependency of reference genome need to be overcome, and the SMART platform may be a transitional technology. Application of high throughput NGS to whole genome sequencing for higher eukaryotes also needs to be introduced as soon as possible, for achievement of the dream, \$1000 per genome. Similarly, improved BLAST or sequence comparison tools and genome informatics pipeline need attention for error free annotation, data repository and retrieval, single nucleotide based analysis, and various other applications in biomedical, evolutionary, and genome wide studies. More structured and accurate data availability is also important. Although manual curation and annotation is highly recommended, but due to increased availability of raw genome information in current days, automation and preferably, NLP based approaches of annotation could be useful in addressing

quality control issues associated with rapid annotation. Visualization and genome mapping tools demand less complexity and better representability. The future of comparative genomics will depend on how fast we can overcome the discussed limitations. While technology and informatics are concerned, we need NGS with longer reads and assembly must be automated, preferably without a reference genome. The technology also demands high speed and accuracy. The mysterious behaviors of most of the microbes are hidden in hypothetical genes or accessory genes. Therefore, more attention is required to address functionality of such genes using various comparative, functional, and structural genomics approaches. The applications of comparative genomics in bacteria are mostly identification of species, genus, strains, phylogenetics, GC rich or AT rich genomes, pan, core, dispensable, in dispensable genomes, PAIs, virulence factors, toxins, drug and vaccine targets, among others. So far we have sequenced 15 *C. pseudotuberculosis* strains and subsequently, genome analysis demonstrates that the pathogen can be regarded as a model organism for the species. *Corynebacterium*. *C. diphtheriae*, *C. glutamicum*, *C. efficiens*, and *C. ulcerans* have been studied to a certain extent at genome level. Nevertheless, our extensive genome sequencing of *C. pseudotuberculosis* strains and subsequent comparative, pan-genome, and subtractive genomics based studies have revealed many hidden characteristics of the genus *Corynebacterium*. Further exploration of *C. pseudotuberculosis* genome informatics will throw more insights and better understanding of the genus, and its various aspects.

Acknowledgments

CNPq and FAPEMIG;

TWAS-CNPq for PhD Fellowship.

References

1. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269: 496-512.

2. Richardson EJ, Watson M (2013) The automatic annotation of bacterial genomes. *Brief Bioinform* 14: 1-12.
3. Markowitz VM (2007) Microbial genome data resources. *Curr Opin Biotechnol* 18: 267-272.
4. Pagani I, Liolios K, Jansson J, Chen IM, Smirnova T, et al. (2012) The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* 40: D571-D579.
5. Morozova O, Marra MA (2008) Applications of next-generation sequencing technologies in functional genomics. *Genomics* 92: 255-264.
6. MacLean D, Jones JD, Studholme DJ (2009) Application of 'next-generation' sequencing technologies to microbial genetics. *Nat Rev Microbiol* 7: 287-296.
7. Metzker ML (2010) Sequencing technologies-the next generation. *Nat Rev Genet* 11: 31-46.
8. Richardson EJ, Watson M (2013) The automatic annotation of bacterial genomes. *Brief Bioinform* 14: 1-12.
9. Romualdi A, Siddiqui R, Glöckner G, Lehmann R, Sühnel J (2005) GenColors: accelerated comparative analysis and annotation of prokaryotic genomes at various stages of completeness. *Bioinformatics* 21: 3669-3671.
10. Lima T, Auchincloss AH, Coudert E, Keller G, Michoud K, et al. (2009) HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. *Nucleic Acids Res* 37: D471-D478.
11. Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, et al. (2012) IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic Acids Res* 40: D115-D222.
12. D'Afonseca, Soares V, Ali SC, Santos A, Pinto AR, et al. (2012) Reannotation of the *Corynebacterium diphtheriae* NCTC13129 genome as a new approach to studying gene targets connected to virulence and pathogenicity in diphtheria.
13. Relman DA (2011) Microbial genomics and infectious diseases. *N Engl J Med* 365: 347-357.
14. Venton D (2012) Highlight: tiny bacterial genome opens a huge mystery: AT mutational bias in *Hodgkinia*. *Genome Biol Evol* 4: 28-29.
15. Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, et al. (2008) The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J Bacteriol* 190: 6881-6893.
16. Höhl M, Kurtz S, Ohlebusch E (2002) Efficient multiple genome alignment. *Bioinformatics* 18: S312-S320.
17. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, et al. (2004) Versatile and open software for comparing large genomes. *Genome Biol* 5: R12.
18. Vallenet D, Engelen S, Mornico D, Cruveiller S, Fleury L, et al. (2009) MicroScope: a platform for microbial genome annotation and comparative genomics. *Database (Oxford)* 2009: bap021.
19. Lukjancenko O, Ussery DW, Wassenaar TM (2012) Comparative genomics of *Bifidobacterium*, *Lactobacillus* and related probiotic genera. *Microb Ecol* 63: 651-673.
20. Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, et al. (2008) The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J Bacteriol* 190: 6881-6893.
21. Trost E, Al-Dilaimi A, Papavasiliou P, Schneider J, Viehoveer P, et al. (2011) Comparative analysis of two complete *Corynebacterium ulcerans* genomes and detection of candidate virulence factors. *BMC Genomics* 12: 383.
22. Metzker ML (2005) Emerging technologies in DNA sequencing. *Genome Res* 15: 1767-1776.
23. Zhang J, Chiodini R, Badr A, Zhang G (2011) The impact of next-generation sequencing on genomics. *J Genet Genomics* 38: 95-109.
24. Pareek CS, Smoczynski R, Tretyn A (2011) Sequencing technologies and genome sequencing. *J Appl Genet* 52: 413-435.
25. Metzker ML (2010) Sequencing technologies-the next generation. *Nat Rev Genet* 11: 31-46.
26. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215: 403-410.
27. Flicek P, Birney E (2009) Sense from sequence reads: methods for alignment and assembly. *Nat Methods* 6: S6-S12.
28. Wooley JC, Godzik A, Friedberg I (2010) A primer on metagenomics. *PLoS Comput Biol* 6: e1000667.
29. Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV (2001) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res* 11: 356-372.
30. Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, et al. (1999) Alignment of whole genomes. *Nucleic Acids Res* 27: 2369-2376.
31. Nakato R, Gotoh O (2010) Cgaln: fast and space-efficient whole-genome alignment. *BMC Bioinformatics* 11: 224.
32. Darling AE, Treangen TJ, Messeguer X, Perna NT (2007) Analyzing patterns of microbial evolution using the mauve genome alignment system. *Methods Mol Biol* 396: 135-152.
33. Darling AE, Mau B, Perna NT (2010) ProgressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 5: e11147.
34. Delcher AL, Phillippy A, Carlton J, Salzberg SL (2002) Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res* 30: 2478-2483.
35. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, et al. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res* 13: 721-731.
36. Brudno M (2007) An introduction to the Lagan alignment toolkit. *Methods Mol Biol* 395: 205-220.
37. Dewey CN (2007) Aligning multiple whole genomes with Mercator and MAVID. *Methods Mol Biol* 395: 221-236.
38. Schatz MC, Trapnell C, Delcher AL, Varshney A (2007) High-throughput sequence alignment using Graphics Processing Units. *BMC Bioinformatics* 8: 474.
39. Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV (2001) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res* 11: 356-372.
40. Cerdeira LT, Carneiro AR, Ramos RT, de Almeida SS, D'Afonseca V, et al. (2011) Rapid hybrid de novo assembly of a microbial genome using only short reads: *Corynebacterium pseudotuberculosis* 119 as a case study. *J Microbiol Methods* 86: 218-223.
41. Siezen RJ, van Hijum SA (2010) Genome (re-)annotation and open-source annotation pipelines. *Microb Biotechnol* 3: 362-369.
42. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, et al. (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9: 75.
43. Gibson R, Smith DR (2003) Genome visualization made fast and simple. *Bioinformatics* 19: 1449-1450.
44. Sato N, Ehira S (2003) GenoMap, a circular genome data viewer. *Bioinformatics* 19: 1583-1584.
45. Kerkhoven R, van Enckevort FH, Boekhorst J, Molenaar D, Siezen RJ (2004) Visualization for genomics: the Microbial Genome Viewer. *Bioinformatics* 20: 1812-1814.
46. Stothard P, Wishart DS (2005) Circular genome visualization and exploration using CGView. *Bioinformatics* 21: 537-539.
47. Hallin PF, Ussery DW (2004) CBS Genome Atlas Database: a dynamic storage for bioinformatic results and sequence data. *Bioinformatics* 20: 3682-3686.
48. Stothard P, Van Domselaar G, Shrivastava S, Guo A, O'Neill B, et al. (2005) BacMap: an interactive picture atlas of annotated bacterial genomes. *Nucleic Acids Res* 33: D317-320.
49. Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J (2009) DNAPlotter: circular and linear interactive genome visualization. *Bioinformatics* 25: 119-120.
50. Alikhan NF, Petty NK, Ben Zakour NL, Beatson SA (2011) BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 12: 402.
51. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, et al. (2009) Circos: an

- information aesthetic for comparative genomics. Genome Res 19: 1639-1645.
52. Grant JR, Arantes AS, Stothard P (2012) Comparing thousands of circular genomes using the CGView Comparison Tool. BMC Genomics 13: 202.
 53. Barh D, Jain N, Tiwari S, Parida BP, D'Afonseca V, et al. (2011) A novel comparative genomics analysis for common drug and vaccine targets in *Corynebacterium pseudotuberculosis* and other CMN group of human pathogens. Chem Biol Drug Des 78: 73-84.
 54. Ott L, McKenzie A, Baltazar MT, Britting S, Bischof A, et al. (2012) Evaluation of invertebrate infection models for pathogenic *Corynebacteria*. FEMS Immunol Med Microbiol 65: 413-421.
 55. Ilatovskiy A, Petukhov M (2009) Genome-wide search for local DNA segments with anomalous GC-content. J Comput Biol 16: 555-564.
 56. Nakabachi A, Yamashita A, Toh H, Ishikawa H, Dunbar HE, et al. (2006) The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. Science 314: 267.
 57. Van Leuven JT, McCutcheon JP (2012) An AT mutational bias in the tiny GC-rich endosymbiont genome of *Hodgkinia*. Genome Biol Evol 4: 24-27.
 58. Pallen MJ, Wren BW (2007) Bacterial pathogenomics. Nature 449: 835-842.
 59. Lightfield J, Fram NR, Ely B (2011) Across bacterial phyla, distantly-related genomes with similar genomic GC content have similar patterns of amino acid usage. PLoS One 6: e17677.
 60. Nishida H (2012) Genome DNA sequence variation, evolution, and function in bacteria and *Archaea*. Curr Issues Mol Biol 15: 19-24.
 61. Röttger R, Kalaghatgi P, Sun P, Soares Sde C, Azevedo V, et al. (2013) Density parameter estimation for finding clusters of homologous proteins--tracing actinobacterial pathogenicity lifestyles. Bioinformatics 29: 215-222.
 62. Ruiz JC, D'Afonseca V, Silva A, Ali A, Pinto AC, et al. (2011) Evidence for reductive genome evolution and lateral acquisition of virulence functions in two *Corynebacterium pseudotuberculosis* strains. PLoS One 6: e18551.
 63. Williamson LH (2001) Caseous lymphadenitis in small ruminants. Vet Clin North Am Food Anim Pract 17: 359-371.
 64. Dorella FA, Pacheco LG, Oliveira SC, Miyoshi A, Azevedo V (2006) *Corynebacterium pseudotuberculosis*: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. Vet Res 37: 201-218.
 65. Baumbach J, Tauch A, Rahmann S (2009) Towards the integrated analysis, visualization and reconstruction of microbial gene regulatory networks. Brief Bioinform 10: 75-83.
 66. Sharan R, Ulitsky I, Shamir R (2007) Network-based prediction of protein function. Mol Syst Biol 3: 88.
 67. Levy ED, Pereira-Leal JB (2008) Evolution and dynamics of protein interactions and networks. Curr Opin Struct Biol 18: 349-357.
 68. Barh D, Gupta K, Jain N, Khatri G, León-Sicairens N, et al. (2013) Conserved host-pathogen PPIs. Globally conserved inter-species bacterial PPIs based conserved host-pathogen interactome derived novel target in *C. pseudotuberculosis*, *C. diphtheriae*, *M. tuberculosis*, *C. ulcerans*, *Y. pestis*, and *E. coli* targeted by Piper betel compounds. Integr Biol (Camb) 5: 495-509.
 69. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 96: 4285-4288.
 70. Enright AJ, Iliopoulos I, Kyripides NC, Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. Nature 402: 86-90.
 71. Dandekar T, Snel B, Huynen M, Bork P (1998) Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci 23: 324-328.
 72. Röttger R, Rückert U, Taubert J, Baumbach J (2012) How little do we actually know? On the size of gene regulatory networks. IEEE/ACM Trans Comput Biol Bioinform 9: 1293-1300.
 73. Baumbach J, Rahmann S, Tauch A (2009) Reliable transfer of transcriptional gene regulatory networks between taxonomically related organisms. BMC Syst Biol 3: 8.
 74. Tetz VV (2005) The pangenome concept: a unifying view of genetic information. Med Sci Monit 11: HY24-HY29.
 75. Medini D, Donati C, Tettelin H, Maignani V, Rappuoli R (2005) The microbial pan-genome. Curr Opin Genet Dev 15: 589-594.
 76. Ventura M, Canchaya C, Tauch A, Chandra G, Fitzgerald GF, et al. (2007) Genomics of Actinobacteria: tracing the evolutionary history of an ancient phylum. Microbiol Mol Biol Rev 71: 495-548.
 77. Soares SC, Silva A, Trost E, Blom J, Ramos R, et al. (2013) The Pan-genome of the animal pathogen *Corynebacterium pseudotuberculosis* Reveals differences in genome plasticity between the biovar ovis and equi strains. PLoS One 8: e53818.
 78. Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, et al. (2008) The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. J Bacteriol 190: 6881-6893.
 79. Reno ML, Held NL, Fields CJ, Burke PV, Whitaker RJ (2009) Biogeography of the *Sulfolobus islandicus* pan-genome. Proc Natl Acad Sci U S A 106: 8605-8610.
 80. Trost E, Blom J, Soares Sde C, Huang IH, Al-Dilaimi A, et al. (2012) Pangenomic study of *Corynebacterium diphtheriae* that provides insights into the genomic diversity of pathogenic isolates from cases of classical diphtheria, endocarditis, and pneumonia. J Bacteriol 194: 3199-3215.
 81. Maurelli AT, Fernández RE, Bloch CA, Rode CK, Fasano A (1998) "Black holes" and bacterial pathogenicity: a large genomic deletion that enhances the virulence of *Shigella* spp. and enteroinvasive *Escherichia coli*. Proc Natl Acad Sci U S A 95: 3943-3948.
 82. Schmidt H, Hensel M (2004) Pathogenicity islands in bacterial pathogenesis. Clin Microbiol Rev 17: 14-56.
 83. Hacker J, Carniel E (2001) Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. EMBO Rep 2: 376-381.
 84. Barcellos FG, Menna P, da Silva Batista JS, Hungria M (2007) Evidence of horizontal transfer of symbiotic genes from a *Bradyrhizobium japonicum* inoculant strain to indigenous diazotrophs *Sinorhizobium (Ensifer) fredii* and *Bradyrhizobium elkanii* in a Brazilian Savannah soil. Applied and environmental microbiology 73: 2635-2643.
 85. Krizova L, Nemec A (2010) A 63 kb genomic resistance island found in a multidrug-resistant *Acinetobacter baumannii* isolate of European clone I from 1977. J Antimicrob Chemother 65: 1915-1918.
 86. Tumapa S, Holden MT, Vesaratchavest M, Wuthiekanun V, Limmathurtsakul D, et al. (2008) *Burkholderia pseudomallei* genome plasticity associated with genomic island variation. BMC Genomics 9: 190.
 87. D'Auria G, Jiménez-Hernández N, Peris-Bondia F, Moya A, Latorre A (2010) *Legionella pneumophila* pangenome reveals strain-specific virulence factors. BMC Genomics 11: 181.
 88. Kittichotirat W, Bumgarner RE, Asikainen S, Chen C (2011) Identification of the pangenome and its components in 14 distinct *Aggregatibacter actinomycetemcomitans* strains by comparative genomic analysis. PLoS One 6: e22420.
 89. Mira A, Martín-Cuadrado AB, D'Auria G, Rodríguez-Valera F (2010) The bacterial pan-genome: a new paradigm in microbiology. Int Microbiol 13: 45-57.
 90. Tauch A, Schneider J, Szczepanowski R, Tilker A, Viehoveer P, et al. (2008) Ultrafast pyrosequencing of *Corynebacterium kroppenstedtii* DSM44385 revealed insights into the physiology of a lipophilic *Corynebacterium* that lacks mycolic acids. J Biotechnol 136: 22-30.
 91. Dobrindt U, Janke B, Piechaczek K, Nagy G, Ziebuhr W, et al. (2000) Toxin genes on pathogenicity islands: impact for microbial evolution. Int J Med Microbiol 290: 307-311.
 92. Soares SC, Abreu VA, Ramos RT, Cerdeira L, Silva A, et al. (2012) PIPS: pathogenicity island prediction software. PLoS One 7: e30848.
 93. Buck GA, Cross RE, Wong TP, Loera J, Groman N (1985) DNA relationships among some tox-bearing *Corynebacteriophages*. Infect Immun 49: 679-684.
 94. Cerdeño-Tarraga AM, Efstratiou A, Dover LG, Holden MT, Pallen M, et al. (2003) The complete genome sequence and analysis of *Corynebacterium diphtheriae* NCTC13129. Nucleic Acids Res 31: 6516-6523.
 95. Groman N, Schiller J, Russell J (1984) *Corynebacterium ulcerans* and

- Corynebacterium pseudotuberculosis* responses to DNA probes derived from corynephage beta and *Corynebacterium diphtheriae*. Infect Immun 45: 511-517.
96. Raskin DM, Seshadri R, Pukatzki SU, Mekalanos JJ (2006) Bacterial genomics and pathogen evolution. Cell 124: 703-714.
 97. Snipen L, Ussery DW (2010) Standard operating procedure for computing pangenome trees. Stand Genomic Sci 2: 135-141.
 98. Urwin R, Maiden MC (2003) Multi-locus sequence typing: a tool for global epidemiology. Trends Microbiol 11: 479-487.
 99. Bolt F (2009) The population structure of the *Corynebacterium diphtheriae* group.
 100. Bolt F, Cassiday P, Tondella ML, Dezoysa A, Efstratiou A, et al. (2010) Multilocus sequence typing identifies evidence for recombination and two distinct lineages of *Corynebacterium diphtheriae*. J Clin Microbiol 48: 4177-4185.
 101. Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, et al. (2012) Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. Microbiology 158: 1005-1015.
 102. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 35: W182-185.
 103. Koski LB, Gray MW, Lang BF, Burger G (2005) AutoFACT: an automatic functional annotation and classification tool. BMC Bioinformatics 6: 151.
 104. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, et al. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11: 119.

Citation: Ali A, Soares SC, Barbosa E, Santos AR, Barh D, et al. (2013) Microbial Comparative Genomics: An Overview of Tools and Insights Into The Genus *Corynebacterium*. J Bacteriol Parasitol 4: 167. doi:[10.4172/2155-9597.1000167](https://doi.org/10.4172/2155-9597.1000167)

Submit your next manuscript and get advantages of OMICS Group submissions

Unique features:

- User friendly/feasible website-translation of your paper to 50 world's leading languages
- Audio Version of published paper
- Digital articles to share and explore

Special features:

- 250 Open Access Journals
- 20,000 editorial team
- 21 days rapid review process
- Quality and quick editorial, review and publication processing
- Indexing at PubMed (partial), Scopus, DOAJ, EBSCO, Index Copernicus and Google Scholar etc
- Sharing Option: Social Networking Enabled
- Authors, Reviewers and Editors rewarded with online Scientific Credits
- Better discount for your subsequent articles

Submit your manuscript at: <http://www.omicsonline.org/submission>



1.2 Conclusion

The review brings the systematic narrative of genomic science, strategies, development and applications of microbial comparative genomics. The review also explains the tools, methods and possible challenges in treating genomic data, annotation and particularly comparative genomic and pathogenomic analysis. Majority (if not all) of the reviewed tools, databases and techniques are user-friendly. The genus *Corynebacterium* is selected as model genus for this study due to the fact that its members are well studied. We believe that the comparative genomic approaches discussed in this manuscript will help researcher to design their projects, particularly those with limited computational skills. The tools and techniques applied in this study are considerably straightforward to use and implement and can be applied on larger scale across bacterial genera. We are also confident that the data generated during such studies can assist further research particularly in diagnosis, antibiotics and vaccines development against pathogens.

CHAPTER 2

CAMPYLOBACTER

(C. fetus subspecies)

2.1 Research Article

Campylobacter fetus subspecies: Comparative genomics and prediction of potential virulence targets. *GENE* 2012.

In this article we presented comparative genomic analysis of the genus *Campylobacter*, including pathogens associated with wide range of diseases targeting both animals and human. Among the species, *C. fetus* subspecies *venerealis* NCTC 10354^T and *C. fetus* subspecies *fetus* 82-40 are responsible for serious diseases in animals such as bovine genital campylobacteriosis, infertility, abortions and specifically gastroenteritis in human. The available fifteen complete genomes of *Campylobacter* species on genbank including *Campylobacter fetus* subspecies were analysed. With guidance of Dr. David Ussery and assistance from in-house tools on server at Center for Biological Sequence Analysis, Technical University of Denmark, the pangenomic approach was applied to the both *C. fetus* subspecies and genus as a whole to estimate the conserved core and species pangenome. The conserved core was translated and characterized for their subcellular localization and functional features. Our in-house tool Pathogenicity Islands Prediction Software (PIPS) was used to identify unique candidate regions (PAIs) in the genomes, and pathogenomic strategies were applied to locate potential virulence factors in *C. fetus* subspecies. Finally, the predicted virulence targets and vaccine candidates are analysed for their role in pathogenicity and potential vaccine candidates are predicted and discussed.



Campylobacter fetus subspecies: Comparative genomics and prediction of potential virulence targets

Amjad Ali^a, Siomar C. Soares^a, Anderson R. Santos^a, Luis C. Guimarães^a, Eudes Barbosa^a, Sintia S. Almeida^a, Vinícius A.C. Abreu^a, Adriana R. Carneiro^b, Rommel T.J. Ramos^b, Syeda M. Bakhtiar^a, Syed S. Hassan^a, David W. Ussery^c, Stephen On^d, Artur Silva^b, Maria P. Schneider^b, Andrey P. Lage^e, Anderson Miyoshi^a, Vasco Azevedo^{a,*}

^a Federal University of Minas Gerais, Belo Horizonte, 31907-270, Minas Gerais, Brazil

^b Federal University of Pará, Belém, 66075-110, Pará, Brazil

^c Center for Biological sequence Analysis CBS, Technical University of DK-2800 Kongens Lyngby, Denmark

^d ESR Christchurch Science Centre, PO Box 29-181, Christchurch, New Zealand

^e Laboratório de Bacteriologia Aplicada, Escola de Veterinária, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil

ARTICLE INFO

Article history:

Accepted 30 July 2012

Available online 6 August 2012

Keywords:

Comparative genomics

Campylobacter fetus subspecies

Pan-genomics

Pathogenomics

Virulence factors

Pathogenicity islands

ABSTRACT

The genus *Campylobacter* contains pathogens causing a wide range of diseases, targeting both humans and animals. Among them, the *Campylobacter fetus* subspecies *fetus* and *venerealis* deserve special attention, as they are the etiological agents of human bacterial gastroenteritis and bovine genital campylobacteriosis, respectively. We compare the whole genomes of both subspecies to get insights into genomic architecture, phylogenetic relationships, genome conservation and core virulence factors. Pan-genomic approach was applied to identify the core- and pan-genome for both *C. fetus* subspecies and members of the genus. The *C. fetus* subspecies conserved (76%) proteome were then analyzed for their subcellular localization and protein functions in biological processes. Furthermore, with pathogenomic strategies, unique candidate regions in the genomes and several potential core-virulence factors were identified. The potential candidate factors identified for attenuation and/or subunit vaccine development against *C. fetus* subspecies contain: nucleoside diphosphate kinase (Ndk), type IV secretion systems (T4SS), outer membrane proteins (OMP), substrate binding proteins CjaA and CjaC, surface array proteins, sap gene, and cytolethal distending toxin (CDT). Significantly, many of those genes were found in genomic regions with signals of horizontal gene transfer and, therefore, predicted as putative pathogenicity islands. We found CRISPR loci and *dam* genes in an island specific for *C. fetus* subsp. *fetus*, and T4SS and sap genes in an island specific for *C. fetus* subsp. *venerealis*. The genomic variations and potential core and unique virulence factors characterized in this study would lead to better insight into the species virulence and to more efficient use of the candidates for antibiotic, drug and vaccine development.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

The *Campylobacter* species are characterized by substantial diversity at the taxonomic and pathologic levels (Moolhuijzen et al., 2009). The clinical and economic significance of the genus is well documented, since these bacteria are involved in a wide range of diseases affecting humans and animals (On, 2001; Skirrow, 1994). The spiral-shaped, gram-negative bacteria also targets cattle, swine, and birds; and, it is the main cause of human bacterial gastroenteritis (Harvey and Greenwood, 1983; Skirrow, 1994; Spence et al., 2011). At the species level, *Campylobacter fetus* subspecies *venerealis* (Cfv) are responsible for causing bovine genital campylobacteriosis (BGC, which is a sexually transmissible disease poses a serious economic threat to the meat and dairy industries in Brazil (Stynen et al., 2011), Argentina (Jimenez et al., 2011) Australia and New Zealand (Moolhuijzen et al., 2009)). There are also several studies describing *C. fetus* subspecies *fetus* (Cff),

Abbreviations: Cfv, *Campylobacter fetus* subspecies *venerealis* NCTC 10354^T; Cff, *Campylobacter fetus* subspecies *fetus* 82-40; CRISPR, Clustered Regularly Interspaced Short Palindrome Repeats; T4SS, Type IV secretion systems; PAIs, Pathogenicity islands; PSE, Potentially surface exposed; SEC, Secreted protein.

* Corresponding author. Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Av. Antônio Carlos 6627, Pampulha 31.270-901, Belo Horizonte, Minas Gerais, Brazil. Tel./fax: +55 31 3409 2610.

E-mail addresses: amjad_uni@yahoo.com (A. Ali), siomars@gmail.com (S.C. Soares), anderson2010@gmail.com (A.R. Santos), luisguimaraes.bio@gmail.com (L.C. Guimarães), eudesgvb@gmail.com (E. Barbosa), sintiaalmeida@gmail.com (S.S. Almeida), vinia.abreu@gmail.com (V.A.C. Abreu), carneiroar@gmail.com (A.R. Carneiro), rommelramos@ufpa.br (R.T.J. Ramos), smarriamb@yahoo.co.in (S.M. Bakhtiar), hassan_chemist@yahoo.com (S.S. Hassan), dave@cbs.dtu.dk (D.W. Ussery), Stephen.On@esr.cri.nz (S. On), asilva@ufpa.br (A. Silva), paula@ufpa.br (M.P. Schneider), alage@vet.ufmg.br (A.P. Lage), miyoshi@mono.icb.ufmg.br (A. Miyoshi), vasco@icb.ufmg.br (V. Azevedo). URL: <http://www.icb.ufmg.br> (V. Azevedo).

which causes sporadic abortion in cattle and sheep and enormous economic losses due to sexually transmitted BGC. This indicates the probable higher prevalence in regions where reproduction by natural breeding is customary (Miller and On, 2011; Stynen et al., 2011). One of the major limiting factors in controlling the spread of the disease is, the low specificity of the molecular techniques and other tests used in the diagnosis of BGC, for example, the direct fluorescent-antibody test (On, 2001; Spence et al., 2011). In principle, the *C. fetus* subspecies show different host tissue adaptations; however, at the molecular genetic level, they are essentially indistinguishable (Moolhuijzen et al., 2009). In an attempt to differentiate the subspecies, the PFGE technique was performed, and *C. fetus* subspecies showed 86% similarity based on PFGE-DNA profiles (On and Harrington, 2001). Another way to differentiate the subspecies is with the use of 16S rRNA genes divergence, since these are sensitive to minor mutations, remain hotspots for variations, and are useful evolutionary regulators to estimate the relationships between organisms (Hansson et al., 2008; Raskin et al., 2006). Nevertheless, the two subspecies were found to be nearly identical through DNA–DNA hybridization analysis and 16S rRNA gene sequence divergence (Clarridge, 2004). The findings, altogether, point to the possibility of very few gene targets causing the differentiation between *C. fetus* subspecies. The pan-genome is interesting when the characterization of such species or genus is in question. In principle, low pan-genome diversity could be a sign of a stable environment, in contrast to a high pan-genome variation which could reflect the considerable diversity in species and adaptation to diverse environments (Mira et al., 2010; Snipen and Ussery, 2010). The bacteria belonging to genus *Campylobacter* have small genomes (1.6–2.0 megabases) compared to other genera (Takamiya et al., 2011). For smaller genomes, it is expected that there will be a higher fraction of conserved genes, or core genome, and relatively fewer genes in dispensable parts of the pan-genome. The pan-genome of *Campylobacter* increases with the addition of each new genome; while, on the other hand, the core genome, representing the conserved gene families, remains stable or slightly decreases with each additional genome (Tettelin et al., 2005). The lineage specific/accessory genome contains genes that are unique to the respective organisms and which may also be involved in critical activities of pathogenicity, drug resistance, and stress responses (Ruiz et al., 2011). These factors may increase the adaptability of pathogens to the particular niches they inhabit; however, they are not imperative to the survival of the organism. Moreover, some copies of these genes can be acquired by horizontal gene transfer and accessory genes have also been shown to be over represented in genomic islands (Ogata, et al., 2001; Ruiz et al., 2011). Vaccine development and drug target identification against bacterial pathogenesis remain topics of prime interest to researchers (Rodriguez-Ortega et al., 2006; Wyszynska et al., 2004). Among all the proteins encoded by bacteria, roughly one third, on average, are secreted, surface exposed or potentially surface exposed (PSE). Secreted proteins are believed to be involved in host cell interactions and toxicity (Giombini et al., 2010; Kaakoush et al., 2010). Furthermore, surface proteins have a significant role, being the first in contact with host cells, facilitating adhesion, invasion of the host cells and helping bacteria to prevent host responses (Barinov et al., 2009). Surface proteins could be potential drug targets against bacteria. Nevertheless, potential surface exposed (PSE) proteins are likely to interact with host immune systems and could also be candidates for vaccine development (Lindahl et al., 2005; Pacheco et al., 2011; Rodriguez-Ortega et al., 2006). Similarly, bacteria adjust their metabolism accordingly, in the various host niches, for proliferation and pathogenicity (Samant et al., 2008). Since metabolism is a pre-requisite for survival and virulence, such pathways could be potential targets for anti-microbial therapies (Chakrabarty, 1998). Nucleotide biosynthesis is the key requirement of bacteria for metabolic functions, consequently the corresponding enzyme remain targets for antibiotic, for protection against the bacteria in host (Schlichtman et al., 1995; Sun et al., 2010).

Despite the economic and immediate importance, unfortunately few completely sequenced genomes of *Campylobacter* genus exist in GenBank. With the recent release of complete genome sequence of *C. fetus* subsp. *venerealis* NCTC 10354^T and the availability of taxonomically and molecularly related genome of *C. fetus* subsp. *fetus* 82-40 genome already on NCBI, we aimed to compare the mentioned genomes, in order to characterize potential virulence factors and get insights into the mechanisms of pathogenicity and host specificity of the pathogen.

2. Material and methods

2.1. Genome selection and annotation

The *C. fetus* subsp. *venerealis* NCTC 10354^T, a newly available genome (Stynen et al., 2011) and the *C. fetus* subsp. *fetus* 82-40 complete genome from GenBank were selected, along with: *C. concisus* 13826, *C. curvus* 52592, *C. hominis* ATCC BAA-381, *C. jejuni* RM1221, *C. jejuni* subsp. *doylei* 26997, *C. jejuni* subsp. *jejuni* 81116, *C. jejuni* subsp. *jejuni* 81-176, *C. jejuni* subsp. *jejuni* IA3902, *C. jejuni* subsp. *jejuni* ICDCCJ 07001, *C. jejuni* subsp. *jejuni* M1, *C. jejuni* subsp. *jejuni* NCTC 11168, *C. jejuni* subsp. *jejuni* S3 and *C. lari* RM2100. GenBank files were obtained with their NCBI genome project IDs (GPID),¹ and full-length 16S rRNAs gene sequences were predicted from all genomes using the rRNA gene finder RNAmmer (Lagesen et al., 2007). Table 1 summarizes the basic genomic statistics, accession numbers, and information regarding the source and diseases for selected organisms. Additionally, two *Helicobacter pylori* genomes (*H. pylori* 2017/Accession No. NC_017374.1, and *H. pylori* 52/Accession No. NC_017354.1) were selected as outgroups for the 16S rRNA phylogenetic tree. Gene annotation was extracted from the GenBank files for the selected genomes, except for *Cfv*, as the protein annotation was not available at the time of analysis.

The program Prodigal was used to identify the genes (ORFs) (Hyatt et al., 2010). For control analysis, the number of predicted genes with Prodigal and those of GenBank was cross-checked, to confirm that the numbers of genes are comparable.

2.2. Phylogenetic analysis of ribosomal RNA

The 16S rRNA genes were identified for all fifteen genomes by using the RNAmmer program (Lagesen et al., 2007) along with the two *H. pylori* genomes (*H. pylori* 2017 and *H. pylori* 52), taxonomically close organisms (On, 2001). Among the extracted 16S rRNA gene sequences, those between 1400 and 1700 nucleotides in length and having a score above 1700 predicted by RNAmmer were selected to construct the phylogenetic tree. In the case of multiple 16S rRNA genes identified for a genome, one with a satisfactory RNAmmer score was arbitrarily selected for posterior analysis. The ClustalW (Thompson et al., 1994) program for multiple sequence alignment was used to align the sequences. Program MEGA5 (Tamura et al., 2011) was then used to create the tree with Neighbor-Joining method. Five hundred bootstrap resamplings were done to estimate the consensus tree.

2.3. Conserved genes and gene family definitions

The conserved genes were predicted by BLASTp similarities with default settings between the genomes. The previously used method (Lukjancenko et al., 2012; Tettelin et al., 2005; Zakham et al., 2011), referred to as the 50/50 rule, was followed. According to this criterion, two genes are considered to belong to a single gene family if their amino acid sequences are at least 50% identical over at least 50% of the length of the longest gene. It follows, then, that multiple genes may constitute a single gene family if they follow the 50/50 criterion.

¹ <http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi>.

Table 1

Campylobacter species and subspecies used in comparative analysis. Basic genome statistics, accession numbers, source and disease information of selected *Campylobacter* organisms.

GPID	Genomes	Accession	Length bp	Proteins	% AT	16S rRNAs	Source	Diseases
62521	<i>C. fetus</i> subsp. <i>venerealis</i> NCTC 10354T (<i>Cfv</i>)	AFGH00000000	1,874,244	2092	66	2	Bovine vaginal mucus	Bovine genital campylobacteriosis, abortion
16293	<i>C. fetus</i> subsp. <i>fetus</i> 82-40 (<i>Cff</i>)	NC_008599.1	1,773,615	1719	66	3	Human blood	Infertility, abortions, septicemia
17159	<i>C. concisus</i> 13826	NC_009802.1	2,099,413	2073	60	3	Human feces	Gastroenteritis
17161	<i>C. curvus</i> 52592	NC_009715.1	1,971,264	1947	55	3	Human feces	Gastroenteritis
20083	<i>C. hominis</i> ATCC BAA-381	NC_009714.1	1,714,951	1637	68	3	Human feces	–
303	<i>C. jejuni</i> RM1221	NC_003912.7	1,777,831	1877	69	3	Chicken skin	Food poisoning
17163	<i>C. jejuni</i> subsp. <i>doylei</i> 26997	NC_009707.1	1,845,106	1982	69	3	Human blood	Bacteremia
17953	<i>C. jejuni</i> subsp. <i>jejuni</i> 81116	NC_009839.1	1,628,115	1617	60	3	Human	Food poisoning
16135	<i>C. jejuni</i> subsp. <i>jejuni</i> 81-176	NC_008787.1	1,699,052	1726	69	3	Human	Food poisoning
28907	<i>C. jejuni</i> subsp. <i>jejuni</i> IA3902 ^a	CP001876	1,672,219	1703	69	3	Ovine ^a	Infertility, abortion
47949	<i>C. jejuni</i> subsp. <i>jejuni</i> ICDCCJ 07001	NC_014802.1	1,708,924	1855	69	3	Human	–
38041	<i>C. jejuni</i> subsp. <i>jejuni</i> M1	CP001900	1,616,648	1638	69	3	Human	Gastroenteritis
8	<i>C. jejuni</i> subsp. <i>jejuni</i> NCTC 11168	NC_002163.1	1,641,481	1658	69	3	Human	Food poisoning
45947	<i>C. jejuni</i> subsp. <i>jejuni</i> S3	CP001960	1,724,586	1815	69	3	Chicken	Food poisoning
12517	<i>C. lari</i> RM2100	NC_012039.1	1,571,661	1580	70	3	Human	Gastroenteritis, diarrhea

^a *Campylobacter jejuni* strain associated with sheep abortion.

All genes of a genome are then grouped into gene families. Genes that fail to unite with a family are assigned to their own unique families.

2.4. Proteome comparison and BLAST matrix construction

The gene sequences from all of the selected organisms identified by the Prodigal program (Hyatt et al., 2010) were translated and every gene in every genome (*Campylobacter*) was compared by BLASTp against every other *Campylobacter* gene in the dataset.

The 50/50 rule for constructing conserved gene families, described above, was considered; the genes conforming to the criterion were assembled into the same gene family. The BLAST results were visualized in a BLAST matrix (Binnewies et al., 2005). The matrix generated from the above BLAST hits illustrates the whole genome pair-wise comparison of the protein content of any two genomes, in both percentage of homology and absolute numbers. For visual presentation, the blocks in the matrix which are colored in different intensities indicate the relative levels of homology among different genomes.

2.5. Pan- and core-genome analysis

Results obtained during the BLAST matrix construction were used to generate pan- and core-genome plots, and a pangenome family tree was constructed representing all the genomes (Snipen and Ussery, 2010). Pairs of genes producing reciprocal best hits were considered to be representative of the same gene family. Gene families with at least one gene in common (described above) were plotted in the core genome. The rest of the total, either unmatched or not qualifying according to the criterion, were plotted in the pan genome. The shared genome (core) was calculated for *C. fetus* subspecies (*fetus* and *venerealis*) and for the rest of the total set of genomes (Table 1).

2.6. Prediction of pathogenicity islands

We used PIPS to predict the putative pathogenicity islands (PAIs) of *Cfv* and *Cff*. This is a novel approach based on the detection of multiple PAI features, like: Codon Usage, G+C content deviations, presence of virulence factors, flanking tRNAs, transposase genes and absence in a closely related, non-pathogenic bacterium. PIPS generates a list of putative PAIs and associated files enabling manual curation on the automatically generated data (Soares et al., 2012). To perform those analyses, we predicted putative PAIs *Cfv* and *Cff* using *C. hominis* as a closely related, non-pathogenic bacterium; the predicted data were then analyzed using the Artemis Comparison Tool (ACT) (Carver et al., 2005) for manual curation and validation. In order to observe the plasticity in those regions in comparison to

the other species of genus *Campylobacter*, we plotted all genomes against the reference genome sequences (*Cfv* and *Cff*) using the software BLAST Ring Image Generator (BRIG) (Alikhan et al., 2011).

2.7. Sub-cellular localization prediction of proteins

Sub-cellular localization of *Campylobacter* protein prediction was made by in silico analysis, using the SurfG+ tool. SurfG+ is a pipeline for protein sub-cellular prediction, incorporating commonly used software for motif searches, including SignalP, LipoP and TMHMM, along with novel HMMSEARCH profiles to predict protein retention signals (Barinov et al., 2009). SurfG+ starts by searching for retention signals, lipoproteins, SEC pathway export motifs and transmembrane motifs, generally in this order. If none of these motifs are found in a protein sequence, it is characterized as being cytoplasmic. A novel possibility introduced by SurfG+ is the ability to distinguish between integral membrane proteins versus PSE (potentially surface-exposed). This is done by a parameter that determines the expected cell wall thickness, expressed in amino acids. Using published information or electron microscopy, it is possible to estimate cell wall thickness values for prokaryotic organisms (Barinov et al., 2009; Giombini et al., 2010). *Campylobacter* proteins were classified into four different sub-cellular locations: cytoplasmic, membrane, PSE and secreted. The *Cfv* and *Cff* genomes were compared based on published cell wall thicknesses (8–10 nm) to those of other species of the genus, including *C. concisus*, *C. curvus*, *C. hominis*, *C. lari*, and *C. jejuni* predicted by SurfG+.

2.8. Analysis of core-exoproteins for immunogenic and pathogenic potentials

The core-exoproteome of *C. fetus* subspecies (*Cfv* and *Cff*) was predicted and analyzed for obtaining potential immunogenic proteins. The immunoinformatic pipeline contains: SurfG+ (Barinov et al., 2009) (discussed above), a tool to predict subcellular localization of proteins; TMHMM (Krogh et al., 2001), for prediction of transmembrane helices in proteins; and, NetMHC (Lundegaard et al., 2008a, 2008b), was used for prediction of binding of peptides to different HLA alleles (43 human; 12 non-human, and additional 76 HLA alleles). This time SurfG+ was set to predict only secreted protein (SEC) and potentially surface exposed (PSE) proteins, both groups had peptide intervals removed and mature protein sequences were obtained for further analysis. The draft amino acid sequence was created from each original sequence (SEC and PSE) and subjected to NetMHC and those sequences with strong binding peptides were filtered. BLAST2GO (www.blast2go.org) program was used to analyze and classify the predicted exoproteins for their functional categories, such as molecular and biological processes.

2.9. Virulence factors and vaccine target prediction

Vaxign, a web-based predictor of vaccine candidates, was used to find proteins with immunogenic properties and vaccine development potentials. The pipeline consists of calculations such as sub-cellular localization, transmembrane domain prediction, adhesion, conservation to human and mouse, epitope binding to MHC class I and class II, and protein functional analysis (He et al., 2010). A total of 17443 protein sequence from 10 *Campylobacter* genome have already been analyzed by vaxign (data from website). We used the core genes as input to predict a broad range, vaccine candidates. The *C. fetus* subspecies core genome was subjected to dynamic search tool for prediction of secreted protein (candidates). We gave special attention to the one predicted as secreted, for their importance in pathogenomics and being ideal targets for vaccine development (Maione et al., 2005; Rodriguez-Ortega et al., 2006; Wyszynska et al., 2004). The candidate virulence factors identified, were then subjected to different virulence databases for validation, filtering (Chen et al., 2012; Zhou et al., 2007), and experimental data were sought in the literature.

3. Results and discussion

3.1. Phylogenetic analysis of ribosomal RNA

For bacterial genomes, the 16S rRNA genes are often used for identification of bacteria at the genus and species level; 16S rRNA genes contain variable and conserved regions and, on average, have a length of about 1500 bp (Clarridge, 2004; Lagesen et al., 2007).

The phylogenetic tree, shown in Fig. 1, is based on 16S rRNA gene sequences extracted from 15 genomes, representing 6 *Campylobacter* species and two *H. pylori* (*H. pylori* 2017 and *H. pylori* 52) genomes. Both *C. fetus* subspecies (*Cfv* and *Cff*) are clustered close together based on their 16S rRNA gene sequence homology and due to their greater similarity in genomes. This finding supports previous microbiological and molecular findings for *C. fetus* subspecies (Moolhuijzen et al., 2009). All the *C. jejuni* subspecies are presented by a separated clade, which indicates that these organisms share the same ancestor in their evolutionary events. It has also previously been observed that *C. jejuni* species seem to lack polymorphisms in their 16S rRNA genes; for example, 45 strains of *C. jejuni* and two *C. coli* 16S rRNA gene sequences revealed nine sequence types of the *Campylobacter* strains, with similarities between the different sequence types in the range of 99.6–99.9% (Hansson et al., 2008). On the other hand, the *C. fetus* subspecies positioned close to the *C. hominis* genome, which is located on a separate clade near the *C. curvus* and *C. concisus* which are on the same clade. Among all the *Campylobacter* species, the *C. hominis* species is the one known as non-pathogenic to humans (Lawson et al., 2001). And 16S rDNA and genome statistical (G+C content) analysis indicated that the organism is closer to *Campylobacter* (*C. gracilis* and *C. sputorum*). The *H. pylori* genomes on a distant branch indicate the considerable 16S rRNA gene divergence; however, taxonomically they are quite close to the genus *Campylobacter* (On, 2001). As we observed, the distribution of the species based on 16S rRNA gene sequences (phylogenetic tree) shown in Fig. 1 is comparable to the whole genome comparison and distribution in the BLAST matrix (Fig. 2).

3.2. Proteome comparisons and BLAST matrix

The BLAST matrix is a pair-wise comparison method, in which the number of hits in a given set of proteomes is plotted against each other (Binnewies et al., 2005). The BLAST matrix generated (Material and methods) and the value of each pairwise comparison are given in the box in the matrix (Fig. 2). Whole genome similarity among the fifteen *Campylobacter* genomes (Table 1) has been introduced with inter- and intra-species comparisons. In principle, the intra-species similarities are expected to be greater than the inter-

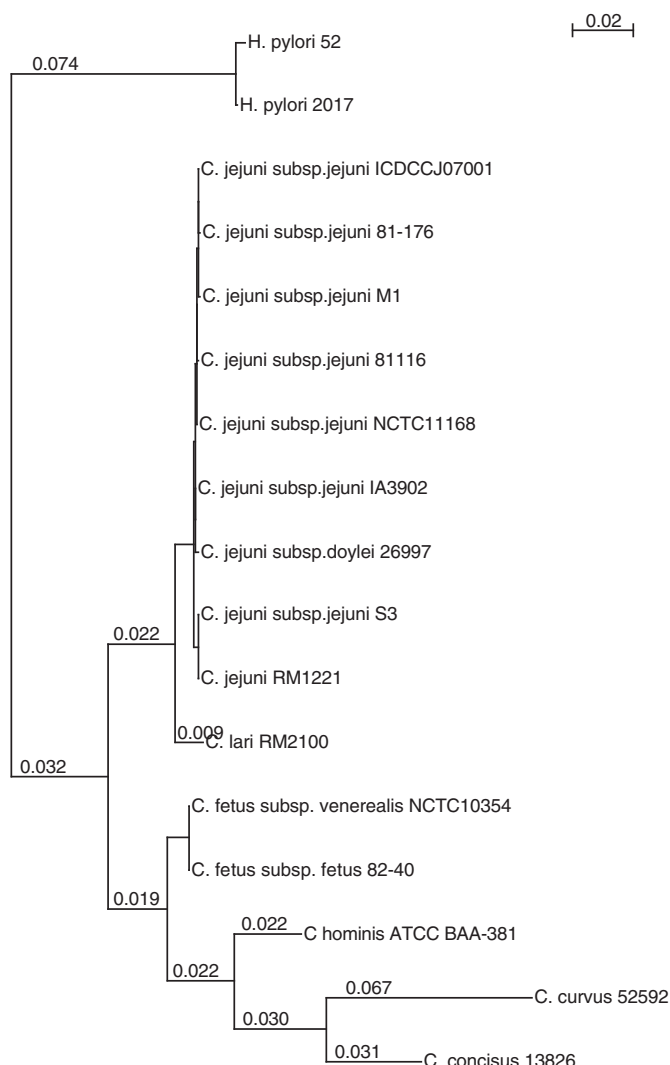


Fig. 1. Phylogenetic analysis of ribosomal RNA. Sequences extracted from 15 *Campylobacter* complete genomes listed in Table 1. Both *Cff* and *Cfv* genomes show greater homology in 16S rRNA gene sequences. The digits indicate the distance (divergence) between species (genomes) in the tree.

species comparisons. We observed the fractions of shared proteins to range from 20% (lowest) to 94% (highest). The lowest fraction (20%) occurred between the genomes *C. jejuni* subsp. *doylei* 26997 and *C. hominis* ATCC BAA 381 due to the fact that, so far, *C. hominis* ATCC BAA 381 is the only *Campylobacter* species that is non-pathogenic to humans (Lawson et al., 2001). The highest genome conservation (94%) was observed between *C. jejuni* subsp. *jejuni* M11 and *C. jejuni* subsp. *jejuni* 81116. Both the subspecies, belong to the *C. jejuni* and have common host and diseases. All of the subspecies of *C. jejuni* species in the matrix have overall greater homology in their proteomes, and the data corresponds to 16S rRNA gene sequences analysis data, where 45 *C. jejuni* species showed 99% similarities (Hansson et al., 2008).

The dense green rectangular near the right corner in the matrix shows *C. fetus* genomes comparison data, both genomes show 76% overall similarity. Where the *Cfv* genome contains a total of 2092 proteins and 2041 protein families, the *Cff* genome contains a total of 1741 proteins and 1705 protein families. The shared proteome of the *C. fetus* subspecies (*Cfv* and *Cff*) consists of 1630 protein families, and the absolute number of protein families goes to 2143. These data correspond to the On and Harrington (2001) analysis of the PFGE-DNA profile for differentiating the subspecies, where the two

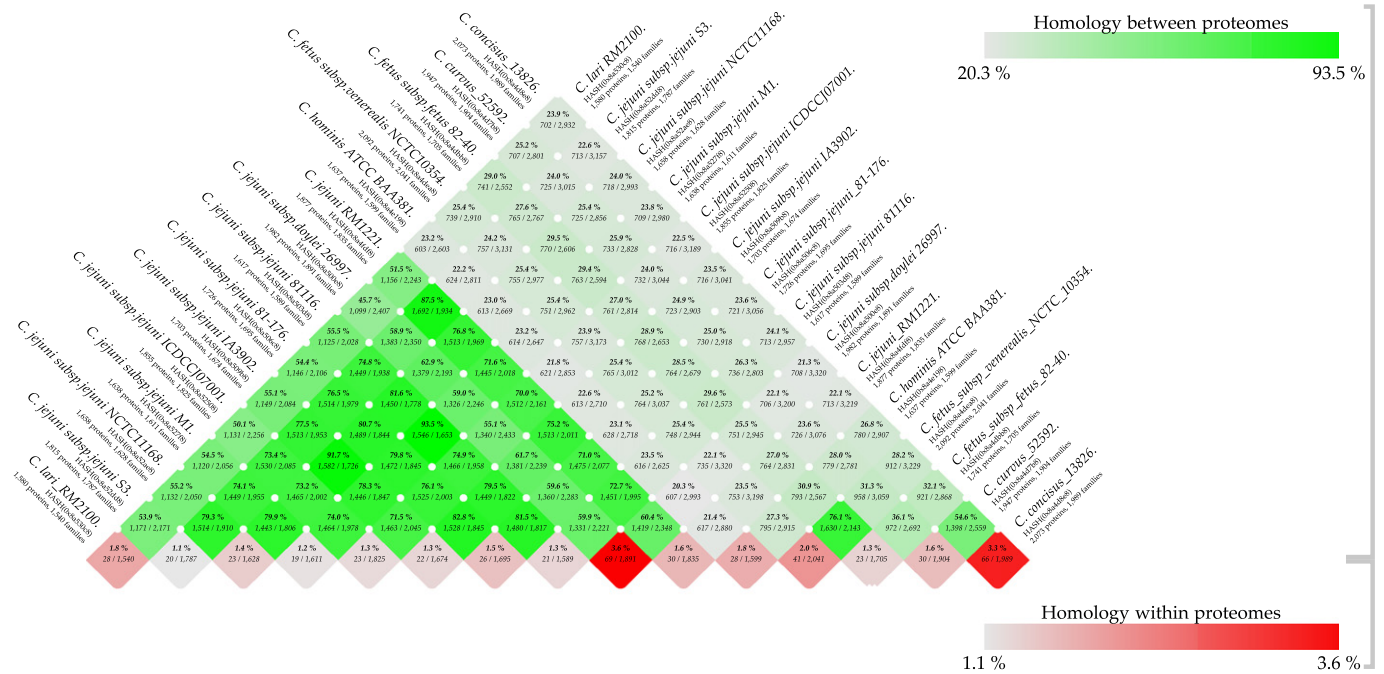


Fig. 2. BLAST matrix: The matrix illustrates the number of preserved gene (protein) families and the total number of gene (protein) families between any two species (genome). The color intensities are based on the relative percentage; the darker the color (green) the greater the conserved gene families; red boxes show the internal homology within the organism's own proteome (scales in percentages). The percentage homology between *C. fetus* subsp. *venerealis* NCTC 10354 and *C. fetus* subsp. *fetus* 82-40 is shown in the corresponding box (76.1%). The dense green pyramid on the left indicates greater homology in *C. jejuni* species and subspecies.

C. fetus subspecies were found to be 86% similar. The overall data in the matrix (genome distribution) also correspond to the 16S ribosomal RNA gene divergence pattern and the phylogenetic tree (Fig. 1).

3.3. Pan- and core-genome plot and analysis

Campylobacter species have small genomes compared to other species, and thus are expected to contain a larger fraction of genes in the core genome and relatively fewer in the dispensable accessory genome (Hepworth et al., 2011). The pan- and core-genome plot was constructed (Fig. 3A), which is a cumulative total of all the conserved core genes, as well as the total number of gene families. The criterion (50/50) for gene families was kept the same in this plot as for the BLAST matrix while the distribution of genomes was changed. The pan/core-genome plot started from *C. fetus* subspecies (*Cff* and *Cfv* respectively) followed by closely related (BLAST matrix and phylogenetic tree) genomes: *C. hominis*, *C. curvus* and *C. concisus*. The *C. jejuni* species are then added in the same order as in the BLAST matrix.

The *Campylobacter* core-genome (15 genomes) comprises 552 gene families and the final pan-genome comprises 7059 gene families. The plot shows that the pan-genome (blue line) increases with the addition of a second genome while the core (red line) genome decreases with the addition of the 2nd genome and drops considerably with the addition of the 3rd genomes.

Due to greater conserved gene families between the *C. fetus* genomes, the core-genome remains stable at the 1st and 2nd positions (*Cff* and *Cfv* respectively). With subsequent additions of *C. hominis* species at the 3rd position and *C. curvus* species at the 4th position the core-genomes continue to decrease, and remain stable with the subsequent *C. jejuni* or slightly decreases, while the pan-genome continues to increase, but at a relatively slow rate.

The core-genome calculated for the *C. fetus* subspecies (*Cfv* and *Cff*), contains 1628 gene families (Additional File 1) and the pan-genome consists of 2144 gene families. The core-genome of the *C. fetus* subspecies was analyzed (www.blast2go.org) for biological functional categories

and, according to Gene Ontology terms, at the third level of biological classification the core-sequences were assigned to different functions and are shown in Fig. 6A. The pie chart displays a greater number of sequences and their role in primary, cellular, and nitrogen metabolic processes. These are followed by macromolecule, biosynthesis, and small molecule metabolic processes.

In addition to the core- and pan-genome estimates, a cluster of 428 gene families was found unique to *Cfv* and a cluster of 88 gene families was found specific to *Cff*.

It is obvious, that the genome similarity and differences are detectable not only by shared genes (core genome) between and among the genomes, but also by the absence of specific genes in specific genome(s). Therefore, the pangenomic tree has been generated based on the presence or absence of specific gene families across the *Campylobacter* genomes (Lukjancenko et al., 2012; Snipen and Ussery, 2010) (Fig. 3B). Again, *Campylobacter* species show greater conserved gene families between them and are positioned closely on the pan-genome dendrogram with greater branch strength. As expected, the genomes of the same species clustered together, and the *C. jejuni* species and subspecies clustered together or near each other.

Interestingly, the *C. fetus* subspecies lie closer to *C. concisus*, *C. curvus* (same branch), and to *C. hominis* species on the other side compared to *C. jejuni* species.

In addition to the core-genome, the lineage specific accessory genes, that is, the group of genes unrelated to any other genes in *Campylobacter* databases, need particular attention (Ruiz et al., 2011). The group of these genes (428 and 88 found in *Cfv* and *Cff*, respectively) does not show any detectable similarity to other available sequences in this study. Often, accessory genes have been found overrepresented in genomic islands (Hsiao et al., 2005) and can be an important fraction of the accessory component. In a microarray-based analysis of 15 clinical isolates of *H. pylori*, the organism closest to *Campylobacter*, up to 56% of the strain-specific genes were orphans (no detectable homologs), and higher numbers were obtained when comparing different sequenced *Rickettsia* genomes (Ogata et al., 2001).

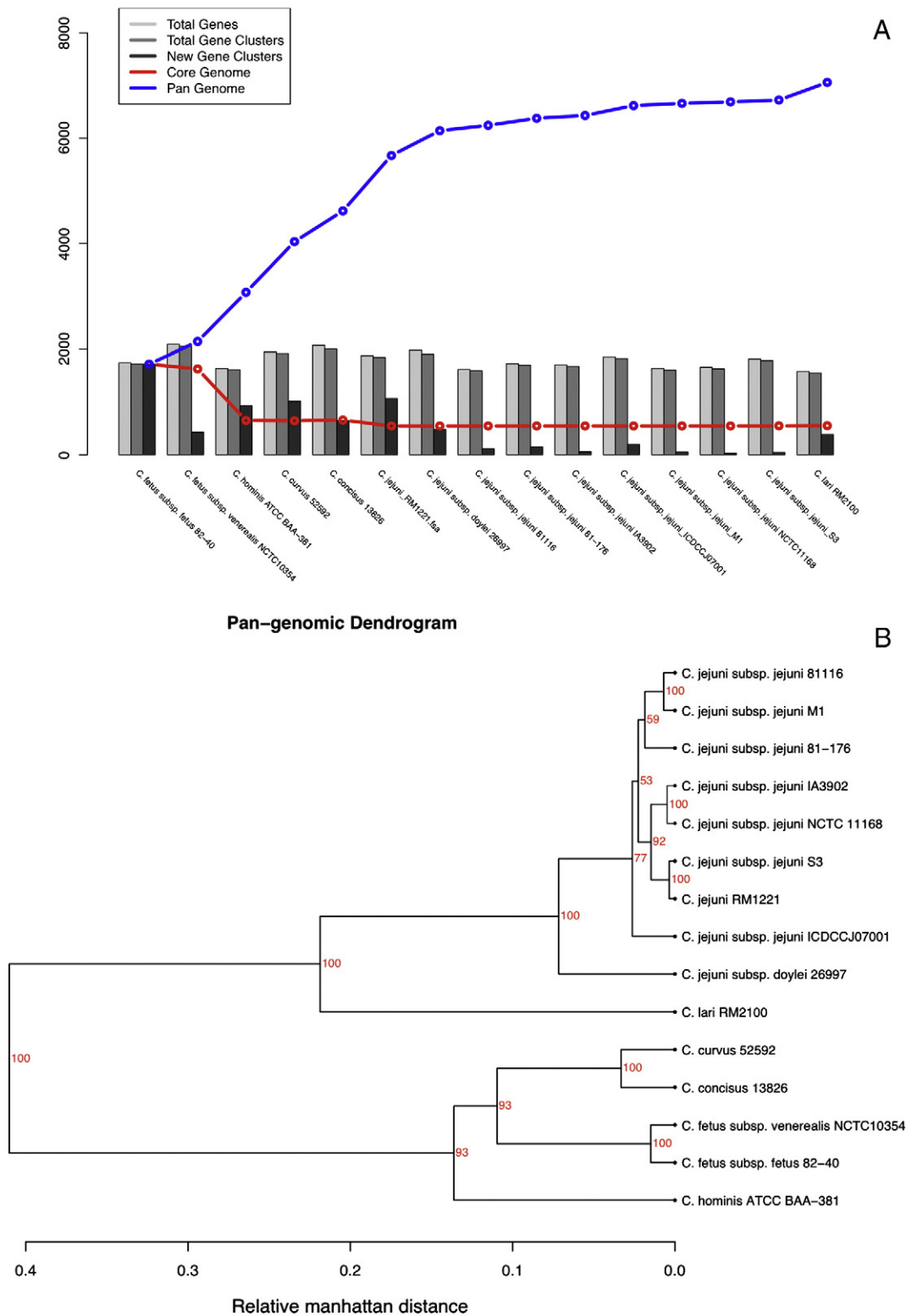


Fig. 3. (A) Pan- and core-genome plot: The line growing towards Y axis indicates the *Campylobacter* pan-genome (7059 gene families) and the line parallel to X-axis indicates the core-genome (552 gene families). The new gene families continue to add to the pool and, hence, the pan is growing while the core-genome remains stable (intra- species) or slightly drops. The *C. fetus* subsp. *fetus* 82-40 and *C. fetus* subsp. *venerealis* NCTC 10354^T genomes are on the 1st and 2nd positions in the row respectively, their core-genome contains 1628 gene families and their pan-genome consists of 2144 gene families. (B) Pangenome family tree. Tree based on the presence or the absence of conserved gene families between *Campylobacter* species (genomes). The relative Manhattan distance indicates how far/close the genomes are located from each other and the digits indicate the strength of the branch.

3.4. Pathogenicity islands in *C. fetus* subspecies

The class of genomic islands which carries virulence factors and, therefore, is responsible for adaptation of pathogenic organisms to

hosts, through horizontal gene transfer, is collectively named pathogenicity islands (PAIs) (Hsiao et al., 2005; Soares et al., 2012).

To access how much of the accessory components and genome plasticity are due to horizontal gene transfer, we have predicted the

putative pathogenicity islands of *Cfv* and *Cff* (hereafter named PICFV and PICFF, respectively). Using the software PIPS [68] we have identified 12 PICFV and 10 PICFF which are represented in Figs. 4A and B. A plot showing the gene organization with genome syntenic breaks in PAIs' regions can be viewed in Additional Fig. 1.

Between those islands, we focused in PICFV5 and 8 and PICFF6 due to big deletion events between *Cfv* and *Cff*, and most prominent gene content.

3.4.1. PICFV5

Accordingly to Fig. 4A, PICFV5 is partially deleted in *Cff*. However, it is worthy of note that a specific region of PICFV5 containing surface array proteins (*sap*) is conserved in both organisms. *sap* genes are responsible for coding surface-layer (S-layer) proteins; capsule-like proteins present on the surfaces of several prokaryotes and having a role in pathogenesis (Thompson, 2002). The S-layer appears attached to bacterial cell surfaces and is, thus, responsible for immune evasion via antiphagocytic properties. Besides, variations in S-layer proteins

occur in vivo, in *C. fetus* species, due to genomic rearrangements within the *sap* genomic locus. These rearrangements are directly linked to antigenic variation; occur in high recombination frequencies among the homologue copies of *sap* genes; and impose a critical problem in the development of vaccines (de Vargas et al., 2002; Thompson, 2002).

3.4.2. PICFV8

Gorkiewicz et al. (2010) through a genomic subtractive-hybridization approach has previously identified a *C. fetus* subsp. *venerealis* specific PAI, which harbors a type IV secretion system (T4SS) along with mobility genes. In a screening for that PAI, it was showed to be present in 51 of 67 *C. fetus* subsp. *venerealis* strains, in contrast with the complete absence in all 45 *C. fetus* subsp. *fetus* strains. Using PIPS we have identified that PAI, herewith denominated PICFV8 (Fig. 4A), with a size of approximately 21 Kb; flanked by a methionyl-tRNA gene relative to *C. fetus* subsp. *fetus*; and disrupting a *putP* gene, which codes for a sodium/proline symporter. T4SS is unique to *Cfv*. The importance of type IV secretion

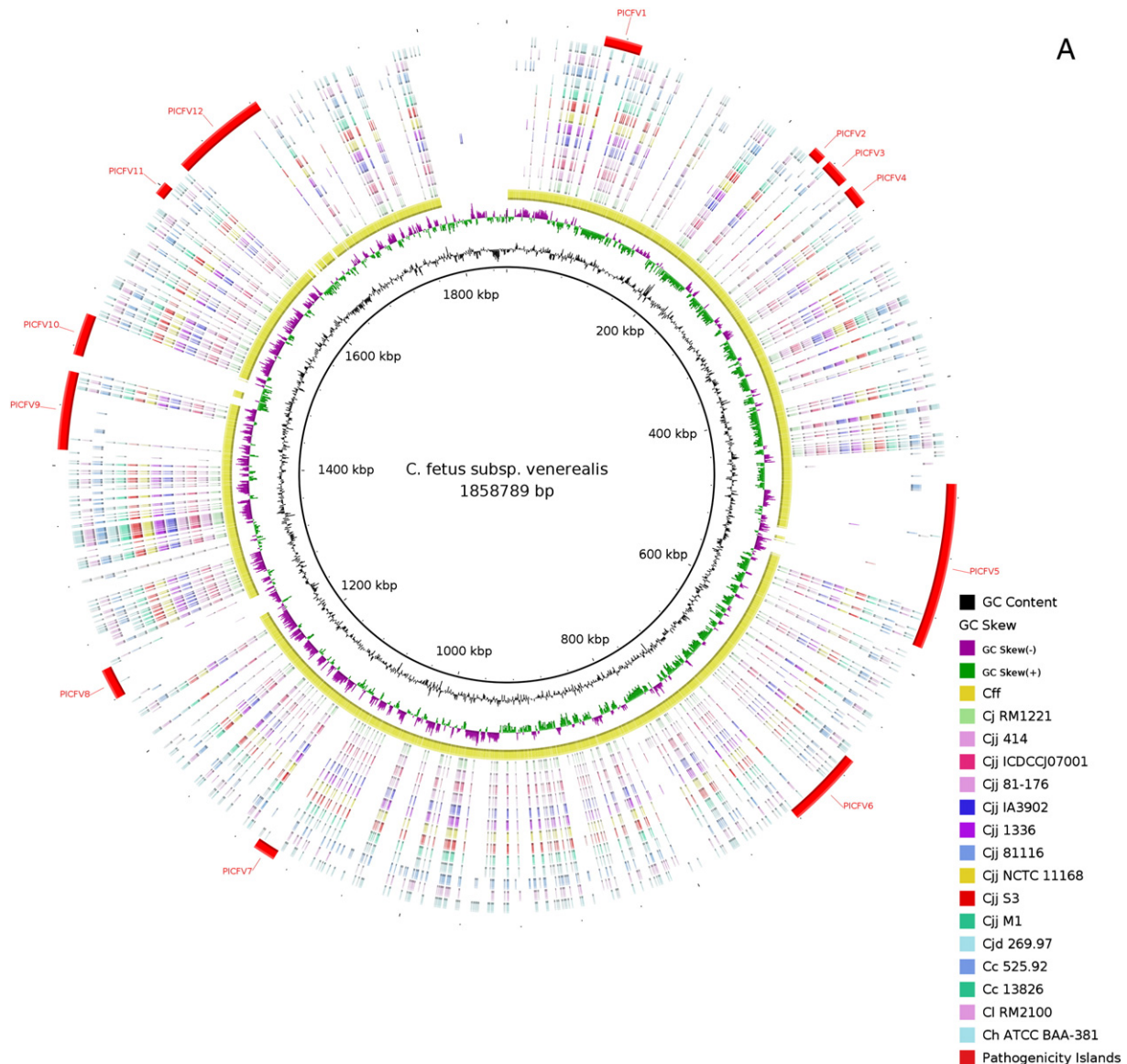


Fig. 4. Genome map comparison of *Campylobacter* species:

- (A) *Campylobacter* species plotted using *C. fetus* subsp. *venerealis* as reference. PICFV, putative pathogenicity island of *C. fetus* subsp. *venerealis*. *Cff*, *C. fetus* subsp. *fetus*; *Cfv*, *C. fetus* subsp. *venerealis*; *Cj*, *C. jejuni*; *Cjj*, *C. jejuni* subsp. *jejuni*; *Cjd*, *C. jejuni* subsp. *doylei*; *Cc* 525.92, *C. curvus* 525.92; *Cc* 13826, *C. concisus* 13826; *Cl*, *C. lari*; *Ch*, *C. hominis*.
 (B) *Campylobacter* species plotted using *C. fetus* subsp. *fetus* as reference. PICFF, putative pathogenicity island of *C. fetus* subsp. *fetus*.

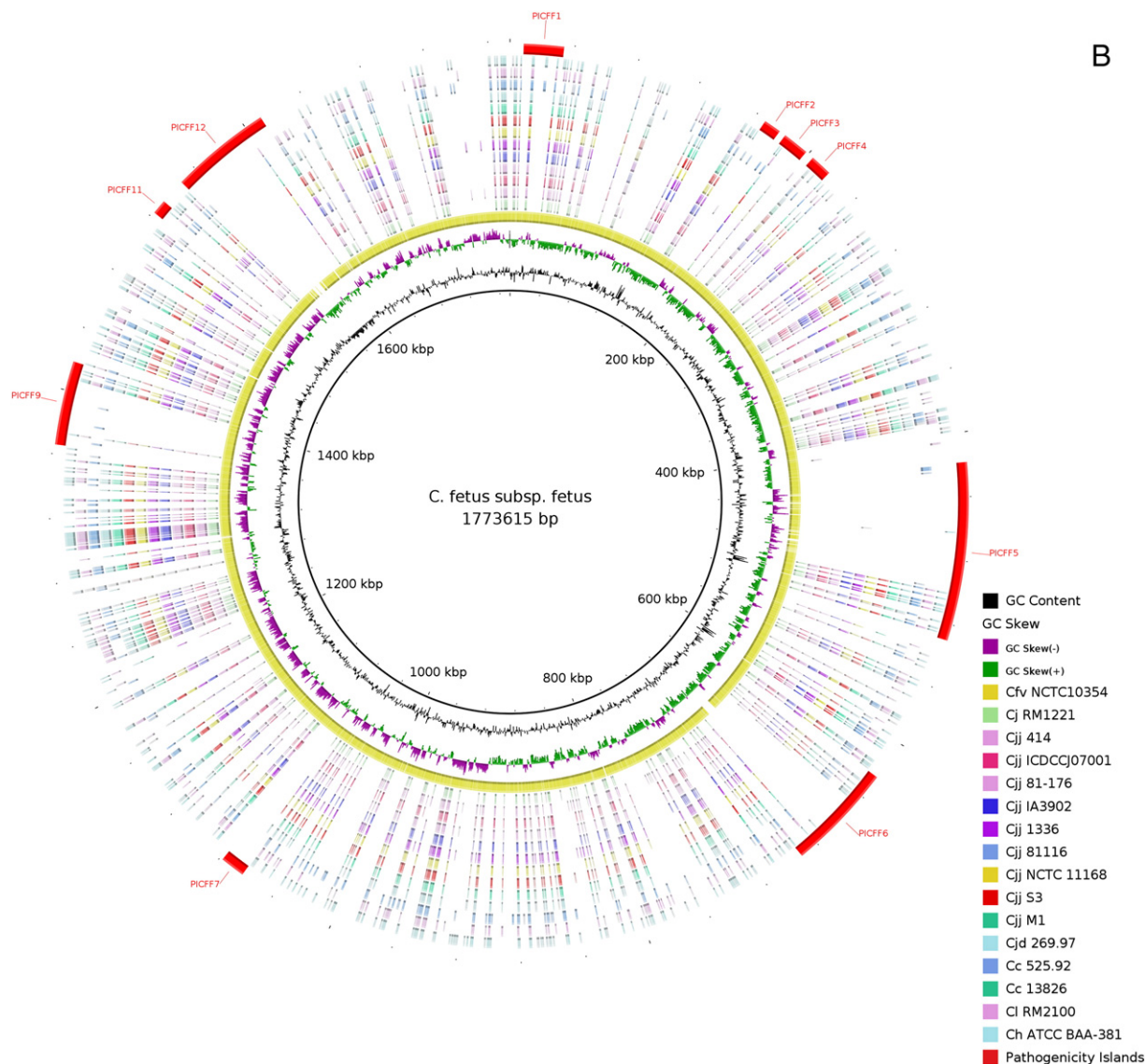


Fig. 4 (continued).

system lies in membrane-associated transport complexes, the transport to distinct target cells and could be an important factor in horizontal gene transfer and possibly be involved in conjugative plasmid transfer or secretion of virulence factors (Cascales and Christie, 2003; Kienesberger et al., 2011; Wallden et al., 2010).

3.4.3. PICFF6

As shown in Fig. 4B, PICFF6 is partially deleted in Cfv and presents Clustered Regularly Interspaced Short Palindrome Repeats (CRISPR) and a dam gene. CRISPRs form a family of short direct repeats, ~25–50 nucleotides, interspaced by unique sequences of similar size. CRISPR

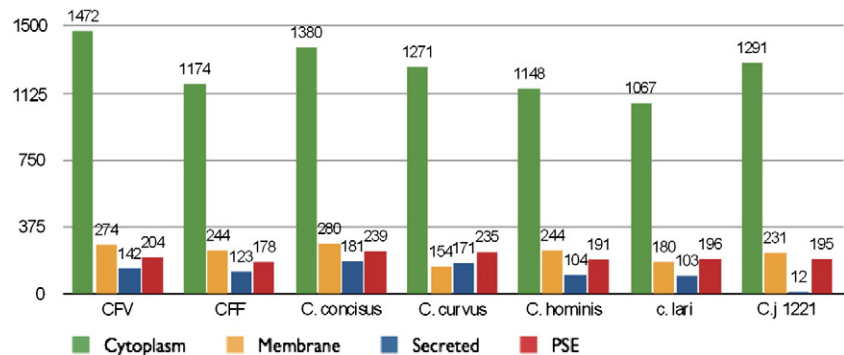


Fig. 5. Subcellular protein location prediction for *Campylobacter* species: Cfv, *C. fetus* subsp. *venerealis* NCTC 10354^T; Cff, *C. fetus* subsp. *fetus* 82-40; Cj1221, *C. jejuni* RM1221; PSE, potential surface exposed. The numbers indicate the proteins in corresponding locations in the cells.

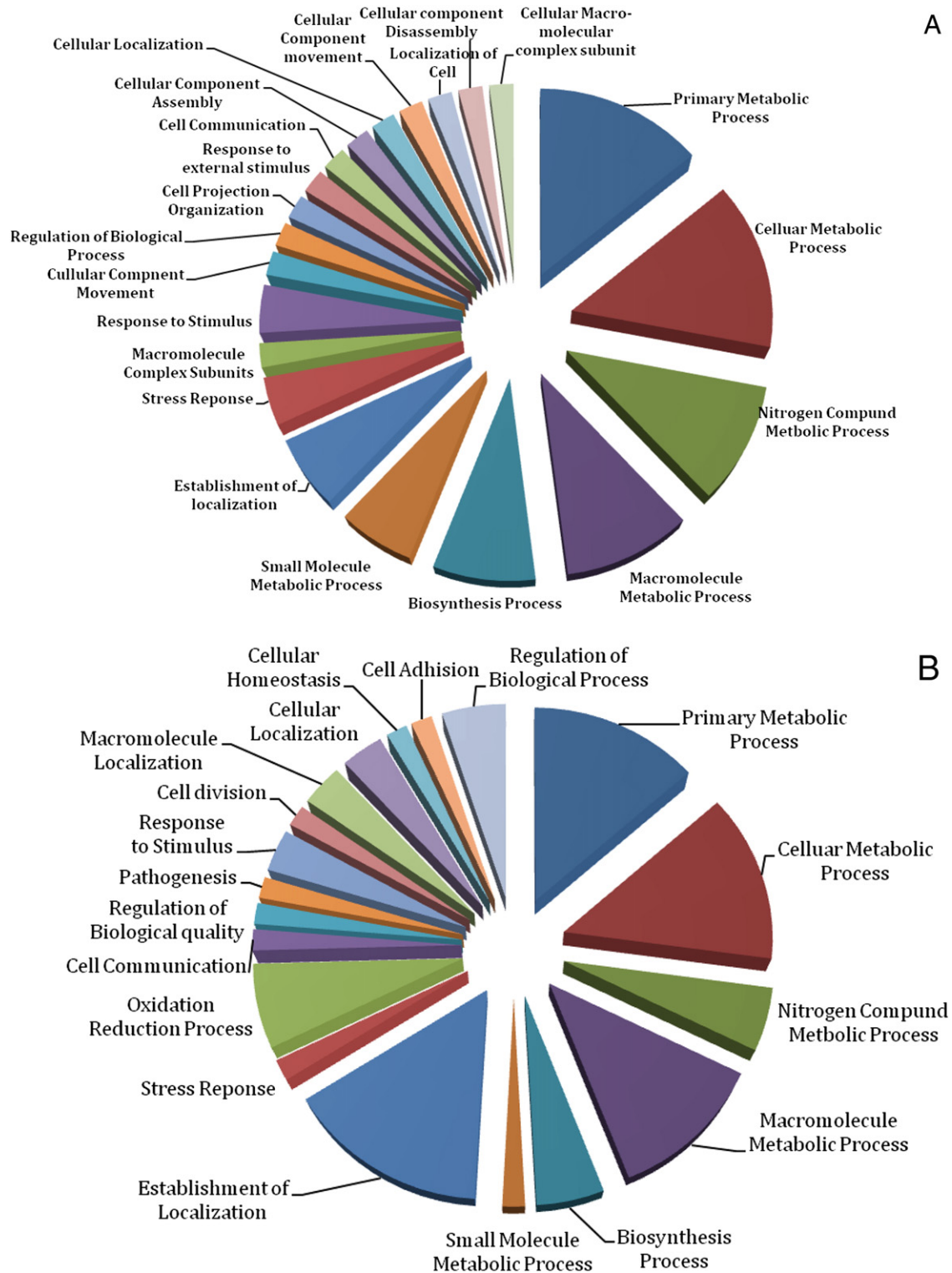


Fig. 6. (A) Chart illustrating the classification of *Campylobacter fetus* subspecies (*Cfv* and *Cff*) core-genome (1628) in the biological process categories. Data generated by BLAST2GO program (www.blast2go.org). (B) Chart illustrating the classification of predicted core-exoproteins (74) for *C. fetus* subspecies (*Cfv* and *Cff*) in the biological process categories. Data generated by BLAST2GO program (www.blast2go.org).

presents similarity with phage and plasmid sequences; they are widespread in prokaryotic organisms; and appear strictly linked with the so called CRISPR associated genes (cas-genes). Cas-genes are involved in a RNA interference type mechanism where the spacers are used as a guide for the cleavage of foreign DNA (Makarova et al., 2006), which consequently, act as a “prokaryotic immune defense

system” against invading foreign replicons (Bolotin et al., 2005; Makarova et al., 2011).

All CRISPRs include *cas1* (metal-dependent DNase) and *cas2* (metal-dependent endoribonuclease) genes, which are involved in the acquisition and integration of the spacers. CRISPR–Cas systems may be divided into three distinct types (I, II or III) accordingly to

the nucleic acid-manipulating activities of the additional cas-genes (*cas3*, *cas9* or *cas10*, respectively) (Makarova et al., 2011).

The *dam* gene codes the Dam enzyme, which catalyzes methylation of adenine residues in GATC sequences through the transfer of the methyl group from S-adenosyl-L-methionine (SAM) using as substrate hemimethylated DNA (Soares et al., 2012). That methylation mechanism plays an important role in restriction–modification (R–M) systems, in which the foreign DNA is recognized and cleaved due to the lack of a specific methylation pattern (Casadesus and Low, 2006; Tu et al., 2003).

Those R–M systems, similar to the CRISPR loci, may have evolved as a form of cellular defense, protecting the cell against foreign DNA (Casadesus and Low, 2006); however, many bacteria lack the cognate restriction enzyme for the methylase enzyme. Those methylases are called “orphan” and may be involved in mismatch repair, initiation of chromosome replication and regulation of gene expression, which, in case of Dam enzyme, is supposed to occur during pre- and post-transcriptional steps (Marinus and Casadesus, 2009). Consequently, the loss of Dam affects several pathways in bacterial physiology and can cause defects, inviability, attenuation and virulence-related defects. Finally, Dam is involved in expression of genes harbored by *Salmonella* pathogenicity island 1 and lack of Dam causes impaired invasion of epithelial cells by Dam[−] *Salmonella* (Lopez-Garrido and Casadesus, 2010). The contents of all predicted pathogenicity islands are given Additional File 4.

3.5. Sub-cellular localization of the *Campylobacter* proteins

Sub-cellular localization of the *Campylobacter* proteins was made by computational analysis using the SurfG+ package tool (Barinov et al., 2009). *C. fetus* subspecies genomes were analyzed for protein localization along with other five species of the genus: *C. concisus*, *C. curvus*, *C. hominis*, *C. lari*, and *C. jejuni*. Fig. 5 displays the numbers of predicted proteins in each sub-cellular location.

Comparative analysis based on subcellular occurrence of the *C. fetus* subspecies (*Cfv* and *Cff*) proteins and other selected *Campylobacter* species proteomes were done with Chi-square tests. The protein ratio

Table 2

Comparative pathogenomics of the genus *Campylobacter*: major/common virulence factors involved in adherence, invasions, motility, secretion systems, and toxins. *Cff*, *C. fetus* subsp. *fetus* 82–40; *Cfv*, *C. fetus* subsp. *venerealis* NCTC 10354[†].

Common virulence factors	<i>Cfv</i>	<i>Cff</i>	Genus <i>Campylobacter</i>
Twitching motility protein	CFV354_1218	CFF8240_1145	Yes
Hemolysin activator-related protein HecB	CFV354_0809	CFF8240_0745	Yes
Phage integrase family protein	CFV354_0922	CFF8240_0843	Yes
Fibronectin-binding protein CadF	CFV354_0255	CFF8240_0194	Yes
Membrane antigen A	CFV354_0532	CFF8240_0472	Yes
Outer membrane protein 18, Omp18	CFV354_1657	CFF8240_1519	Yes
Antigen CjaC	CFV354_1338	CFF8240_1233	Yes
Peptidase U32 (collagenase)	CFV354_1125	CFF8240_1050	Yes
RNase R (VacB homolog)	CFV354_1202	CFF8240_1129	Yes
Invasin InvA	CFV354_1190	CFF8240_1117	Yes
Invasion protein CiaB	CFV354_1385	CFF8240_1280	Yes
Phospholipase PldA	CFV354_0258	CFF8240_0197	Yes
Hsp12 variant C	CFV354_0081	CFF8240_0019	Yes
Type IV secretion system	+	–	No
N-linked protein glycosylation	+	+	Yes
O-linked flagellar glycosylation	+	+	Yes
LOS	–	–	Yes
JlpA	–	–	Yes
PEB1/CBF1	+	+	Yes
Capsule biosynthesis and transport	–	–	Yes
Cytolethal distending toxin (CDT)	CFV354_0087	CFF8240_0026	Yes
Pathogenicity islands	PICFV5–PICFV8	PICFF6	No

of occurrence in different locations (cytoplasmic, membrane anchored, potentially exposed and secreted proteins) was nearly constant in subspecies of *C. fetus* subspecies (*Cfv* and *Cff*) and slight differences in percentage distributions were observed. Although, the total numbers of protein content in both genomes are not close (2092 and 1719 respectively; Table 1). In both *C. fetus* subspecies (*Cfv* and *Cff*), the observed percentage distribution in each sub cellular locations was: cytoplasmic proteins 68% and 70%; membrane proteins 14% and 13%; secreted proteins 7% and 7%; and PSE 10% and 10% respectively. As discussed before, the secreted and potentially surface exposed proteins are considered good candidates, and essential to analyze their immunogenic and pathogenic function (Barinov et al., 2009; Giombini et al., 2010; Rodriguez-Ortega et al., 2006).

3.6. Analysis of core-exoproteins for immunogenic and pathogenic potentials

The importance of secreted and surface exposed proteins in pathogenesis lies in their vital role in interactions with host-cells: adhesion, invasion, toxicity, and protecting bacteria from host cell immune responses. Therefore, they may be good candidates for drug (Lindahl et al., 2005) and vaccine (Maione et al., 2005) development (Pacheco et al., 2011; Rodriguez-Ortega et al., 2006). The translated core-genome (1628 gene families) of the *C. fetus* subspecies was analyzed for prediction of potential candidate proteins, having immunogenic and pathogenic signals (exoproteins; SEC and PSE). We were able to identify a total of 285 sequences, of which 127 were SEC and 158 were PSE. In order to select the statistically best candidate sequences from among the 4 sets of sequences/proteins, the top category (75–100%; data not given) was selected. Following selection, 33 SECs and 44 PSEs were filtered (Additional File 3). The biological and molecular functions of the immunogenic protein sequences were predicted, according to Gene Ontology's functional classification, the orthologous exoprotein sequences, common to the *Cfv* and *Cff*, were classified and are shown in Fig. 6B. As shown in the pie chart, a greater number of sequences have a role in primary, cellular, macromolecule and, nitrogen metabolic processes; these are followed by biosynthesis and small molecule metabolic processes. However, there are proteins which are involved in vital cellular processes, growth, survival and, pathogenesis.

3.7. Virulence targets and vaccine candidates in *C. fetus* subspecies

Seventeen genomic sequences (gene families), potentially virulent, diagnostic and, vaccine targets have been identified. The sequences were then aligned for protein products by BLASTp. Among those identified sequences (gene families), the following were found associated with flagella: flagellar hook associated proteins (FlgK, FlgE, FlgL) (CFF8240_0100, CFF8240_1769, CFF8240_0008, CFF8240_0683, and CFF8240_0092), flagellar basal body protein (CFF8240_0523), and flagellins B protein (CFF8240_1635) (Guerry et al., 1991). The genes for flagellar proteins have been found conserved in *Campylobacter* species, these factors facilitate the colonization and are associated with the processes of chemotaxis and motility (Moolhuijzen et al., 2009). Nevertheless, the importance of flagellins and S-layered proteins are due to the presence of glycoproteins in them and, are well known to be involved in colonization of *Campylobacter* in the host (Thompson, 2002). Furthermore, glycosylation of flagellins is also reported to be important for virulence in animal pathogenic bacteria such as *P. aeruginosa* and *C. jejuni* (Konkel et al., 2001; Yamamoto et al., 2011). The sequence CFF8240_0233, codes for nucleoside diphosphate kinase (Ndk), which is an enzyme involved in nucleoside triphosphate synthesis. Additionally, Ndk is also documented as an important factor in bacterial growth, signal transduction and pathogenicity (Chakrabarty, 1998). Some experimental evidence supports the theory that mycobacterial Ndk is a potential virulence factor, controls phagosome maturation and supports survival of *Mycobacteria* within the macrophage (Sun et al.,

2010). Furthermore, extracellular secretion of NdK has also been previously reported in a number of pathogens including *Pseudomonas aeruginosa*, *Trichinella spiralis*, *Vibrio cholera* and *Mycobacterium bovis* (BCG) (Schlichtman et al., 1995). *Mycobacterium tuberculosis* NdK, was found to be cytotoxic to mouse macrophage cells, in an ATP-dependent P2Z receptor-mediated pathway and is believed to be an important virulence factor. In *Escherichia coli*, NdK was shown to phosphorylate histidine kinases EnvZ and CheA, showing its involvement in signal transduction systems (Kumar et al., 2005). The sequences CFF8240_0484, CFF8240_0464, AY450397.1 and CFF8240_0462 code for surface layer proteins and surface array proteins. These surface proteins provide the primary protection to the pathogen and it is observed that the *C. fetus* outermost crystalline layer is made of monomolecular proteins called S-layer proteins (SLP), where SLP are important factors in providing resistance to host immune defenses and virulence, and thus are considered significant virulence targets (Blaser et al., 1994; Dworkin et al., 1997; Thompson et al., 1998). The *C. fetus* S-layer transporter genes (*sapD*, E and F) are located on an invertible DNA element flanked by two similar but not identical S-layer genes, whose inversion is responsible for the antigenic variation (Thompson et al., 1998). The sequence CFF8240_0215 codes putative DNA-binding/iron metalloprotein/AP endonuclease. Its conserved domains contain glycoprotease, peptidase M22 and O-sialoglycoprotein peptidase. However, its role in biological processes is predicted as proteolysis (GO:0006508) and molecularly acts as peptidase (GO:0008233). Two hypothetical proteins are coded by CFF8240_0738 and CFF8240_1659. The later, however, have been putatively identified as prokaryotic lipoprotein. The sequence CFF8240_0026 encodes cytolethal distending toxin (CDT), an important factor (toxin) that produces resistance. Among other toxins, it is a well known exotoxin in *Campylobacter* (Mooney et al., 2001; Whitehouse et al., 1998). Finally, the sequence CFF8240_0019 encodes the protein Hsp12 variant C. All the predicted gene family sequences (locus_tags) and their GO/annotation are given in Additional File 2.

3.8. Comparative pathogenomics at the genus level

The predicted virulence factors in the *C. fetus* species, along with data from literature searches and virulence databases, were then compared with the genus *Campylobacter*, for their presence and roles in pathogenesis (Chen et al., 2012; Kaakoush et al., 2010). Comparative analysis of the virulence factors, between the *C. fetus* subspecies and other members of the genus, is shown in Table 2. As may be observed, a number of predicted virulence factors were found to be associated with mechanisms of interaction between organism to organism and organism to host. They play major roles in the processes of invasion, adherence, motility, secretion systems and toxins, such as, *InvA*, *CadF*, hemolysins, twitching motility proteins and CDT (Kaakoush et al., 2010; Moolhuijzen et al., 2009). One major virulence factor, the cytolethal distending toxin (CDT), is the exotoxin encoded by genes *cdtA*, *cdtB* and *cdtC*. Among known toxins encoded by *Campylobacter*, the cytolethal distending toxin (CDT) is the one which is fully characterized (Mooney et al., 2001). The *cdt* genes cause cellular detention at the G2 cell cycle and, eventually, death of the cell lines (Whitehouse et al., 1998). Genomes of non *C. jejuni* species containing the *cdt* operon also encode the fibronectin adhesion protein *CadF* (Mooney et al., 2001; Whitehouse et al., 1998). An important membrane protein and potential virulence-associated factor, the outer membrane protein 18 (OMP18), is encoded by *Cff* (CFF8240_1519) and *Cfv* (CFV354_1657). Previous studies report that OMP18 was effective in inducing dendritic cell maturation and function, and was also found initiating a Th-1-mediated immune response (Rathinavelu et al., 2005). The other proteins, for example, the outer membrane substrate-binding proteins *CjaA* and *CjaC* (CFF8240_1233 and CFV354_1338), were found to be involved in amino acid transport: the latter (*CjaC*) being a putative ATP-binding cassette type cysteine transporter, found conserved in some human isolates, and the former (*CjaA*) being required for histidine transport. The *cjaA* gene has been shown to elicit immune response and

protection against wild type *Campylobacter* (Wyszynska et al., 2004). Both *CjaA* and *CjaC* proteins have been proposed as promising candidates for vaccine development against these organisms. Significantly, immunization with a virulent *Salmonella* expressing *Campylobacter cjaA* gene developed serum IgG and mucosal IgA antibody responses against *Campylobacter* membrane proteins and *Salmonella* OMPs, conferring protection in birds (Shoaf-Sweeney et al., 2008; Wyszynska et al., 2004).

4. Conclusion

We were able to document various potential virulence factors, vaccine candidates, and genomic regions associated with *C. fetus* subspecies pathogenicity. The strategy used here may be extended to other bacterial species and genera for the identification of core and unique genes and proteins linked to the vital functions of the organism, such as metabolism, defense mechanisms, and pathogenicity. Furthermore, the data could be incorporated to public pathogenomic databases. However, we would suggest that there is an urgent need for sequencing more *C. fetus* subspecies, to provide better insights into the lifestyle of the organisms.

Author contributions

AA, SCS, ARS, LCG, EB, SSA, VACA, ARC, RTJR and SSH were involved in all analyses for predictions of genes, tRNA, rRNA and genomic analysis. SCS performed the pathogenicity island analysis. ARS predicted the subcellular proteins in the genome. AA was responsible for comparative genomic, pathogenomics and identification of new genetic targets for the development of drugs and vaccines. VA, DWU, AS and AM participated in project design and supervision of the whole project. AA, SO, MPS and SCS, were involved in writing the manuscript. All the authors have read and approved the final manuscript.

Supplementary data to this article can be found online at <http://dx.doi.org/10.1016/j.gene.2012.07.070>.

Acknowledgments

PhD Fellowship by TWAS-CNPq (The Academy of Sciences for Developing World-Conselho Nacional de Desenvolvimento Científico).

References

- Alikhan, N.F., Petty, N.K., Ben Zakour, N.L., Beatson, S.A., 2011. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 12, 402.
- Barinov, A., et al., 2009. Prediction of surface exposed proteins in *Streptococcus pyogenes*, with a potential application to other Gram-positive bacteria. *Proteomics* 9, 61–73.
- Binnewies, T.T., Hallin, P.F., Staerfeldt, H.H., Ussery, D.W., 2005. Genome update: proteome comparisons. *Microbiology* 151, 1–4.
- Blaser, M.J., Wang, E., Tummuru, M.K., Washburn, R., Fujimoto, S., Labigne, A., 1994. High-frequency S-layer protein variation in *Campylobacter fetus* revealed by *sapA* mutagenesis. *Mol. Microbiol.* 14, 453–462.
- Bolotin, A., Quinquis, B., Sorokin, A., Ehrlich, S.D., 2005. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* 151, 2551–2561.
- Carver, T.J., Rutherford, K.M., Berriman, M., Rajandream, M.A., Barrell, B.G., Parkhill, J., 2005. ACT: the Artemis Comparison Tool. *Bioinformatics* 21, 3422–3423.
- Casadesus, J., Low, D., 2006. Epigenetic gene regulation in the bacterial world. *Microbiol. Mol. Biol. Rev.* 70, 830–856.
- Cascales, E., Christie, P.J., 2003. The versatile bacterial type IV secretion systems. *Nat. Rev. Microbiol.* 1, 137–149.
- Chakrabarty, A.M., 1998. Nucleoside diphosphate kinase: role in bacterial growth, virulence, cell signalling and polysaccharide synthesis. *Mol. Microbiol.* 28, 875–882.
- Chen, L., Xiong, Z., Sun, L., Yang, J., Jin, Q., 2012. VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. *Nucleic Acids Res.* 40, D641–D645.
- Claridge III, J.E., 2004. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clin. Microbiol. Rev.* 17, 840–862 (table of contents).

- de Vargas, A.C., Costa, M.M., Vainstein, M.H., Kreutz, L.C., Neves, J.P., 2002. *Campylobacter fetus* subspecies *venerealis* surface array protein from bovine isolates in Brazil. *Curr. Microbiol.* 45, 111–114.
- Dworkin, J., Shedd, O.L., Blaser, M.J., 1997. Nested DNA inversion of *Campylobacter fetus* S-layer genes is recA dependent. *J. Bacteriol.* 179, 7523–7529.
- Giombini, E., Orsini, M., Carrabino, D., Tramontano, A., 2010. An automatic method for identifying surface proteins in bacteria: SLEP. *BMC Bioinforma.* 11, 39.
- Gorkiewicz, G., et al., 2010. A genomic island defines subspecies-specific virulence features of the host-adapted pathogen *Campylobacter fetus* subsp. *venerealis*. *J. Bacteriol.* 192, 502–517.
- Guerry, P., Alm, R.A., Power, M.E., Logan, S.M., Trust, T.J., 1991. Role of two flagellin genes in *Campylobacter* motility. *J. Bacteriol.* 173, 4757–4764.
- Hansson, I., Persson, M., Svensson, L., Engvall, E.O., Johansson, K.E., 2008. Identification of nine sequence types of the 16S rRNA genes of *Campylobacter jejuni* subsp. *jejuni* isolated from broilers. *Acta Vet. Scand.* 50, 10.
- Harvey, S.M., Greenwood, J.R., 1983. Probable *Campylobacter fetus* subsp. *fetus* gastroenteritis. *J. Clin. Microbiol.* 18, 1278–1279.
- He, Y., Xiang, Z., Mobley, H.L., 2010. Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development. *J. Biomed. Biotechnol.* 297–505.
- Hepworth, P.J., et al., 2011. Genomic variations define divergence of water/wildlife-associated *Campylobacter jejuni* niche specialists from common clonal complexes. *Environ. Microbiol.* 13, 1549–1560.
- Hsiao, W.W., Ung, K., Aeschliman, D., Bryan, J., Finlay, B.B., Brinkman, F.S., 2005. Evidence of a large novel gene pool associated with prokaryotic genomic islands. *PLoS Genet.* 1, e62.
- Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W., Hauser, L.J., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinforma.* 11, 119.
- Jimenez, D.F., Perez, A.M., Carpenter, T.E., Martinez, A., 2011. Factors associated with infection by *Campylobacter fetus* in beef herds in the Province of Buenos Aires, Argentina. *Prev. Vet. Med.* 101, 157–162.
- Kaakoush, N.O., et al., 2010. The secretome of *Campylobacter concisus*. *FEBS J.* 277, 1606–1617.
- Kienesberger, S., et al., 2011. Interbacterial macromolecular transfer by the *Campylobacter fetus* subsp. *venerealis* type IV secretion system. *J. Bacteriol.* 193, 744–758.
- Konkel, M.E., Monteville, M.R., Rivera-Amill, V., Joens, L.A., 2001. The pathogenesis of *Campylobacter jejuni*-mediated enteritis. *Curr. Issues Intest. Microbiol.* 2, 55–71.
- Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L., 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305, 567–580.
- Kumar, P., et al., 2005. Nucleoside diphosphate kinase from *Mycobacterium tuberculosis* cleaves single strand DNA within the human c-myc promoter in an enzyme-catalyzed reaction. *Nucleic Acids Res.* 33, 2707–2714.
- Lagesen, K., Hallin, P., Rodland, E.A., Staerfeldt, H.H., Rognes, T., Ussery, D.W., 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* 35, 3100–3108.
- Lawson, A.J., On, S.L., Logan, J.M., Stanley, J., 2001. *Campylobacter hominis* sp. nov., from the human gastrointestinal tract. *Int. J. Syst. Evol. Microbiol.* 51, 651–660.
- Lindahl, G., Stallhammar-Carlemalm, M., Areschoug, T., 2005. Surface proteins of *Streptococcus agalactiae* and related proteins in other bacterial pathogens. *Clin. Microbiol. Rev.* 18, 102–127.
- Lopez-Garrido, J., Casadesus, J., 2010. Regulation of *Salmonella enterica* pathogenicity island 1 by DNA adenine methylation. *Genetics* 184, 637–649.
- Lukjancenko, O., Ussery, D.W., Wassenaar, T.M., 2012. Comparative genomics of *Bifidobacterium*, *Lactobacillus* and related probiotic genera. *Microb. Ecol.* 63, 651–673.
- Lundegaard, C., Lamberth, K., Hamdahl, M., Buus, S., Lund, O., Nielsen, M., 2008a. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8–11. *Nucleic Acids Res.* 36, W509–W512.
- Lundegaard, C., Lund, O., Nielsen, M., 2008b. Accurate approximation method for prediction of class I MHC affinities for peptides of length 8, 10 and 11 using prediction tools trained on 9mers. *Bioinformatics* 24, 1397–1398.
- Maione, D., et al., 2005. Identification of a universal Group B *Streptococcus* vaccine by multiple genome screen. *Science* 309, 148–150.
- Makarova, K.S., Grishin, N.V., Shabalina, S.A., Wolf, Y.I., Koonin, E.V., 2006. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol. Direct* 1, 7.
- Makarova, K.S., Aravind, L., Wolf, Y.I., Koonin, E.V., 2011. Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR–Cas systems. *Biol. Direct* 6, 38.
- Marinus, M.G., Casadesus, J., 2009. Roles of DNA adenine methylation in host–pathogen interactions: mismatch repair, transcriptional regulation, and more. *FEMS Microbiol. Rev.* 33, 488–503.
- Miller, W.G., On, S.L., 2011. International Committee on Systematics of Prokaryotes. Subcommittee on the taxonomy of *Campylobacter* and related bacteria: minutes of the closed meeting, 2 September 2009, Niigata, Japan. *Int. J. Syst. Evol. Microbiol.* 61, 2559–2560.
- Mira, A., Martin-Cuadrado, A.B., D'Auria, G., Rodriguez-Valera, F., 2010. The bacterial pan-genome: a new paradigm in microbiology. *Int. Microbiol.* 13, 45–57.
- Moolhuijzen, P.M., et al., 2009. Genomic analysis of *Campylobacter fetus* subspecies: identification of candidate virulence determinants and diagnostic assay targets. *BMC Microbiol.* 9, 86.
- Mooney, A., Clyne, M., Curran, T., Doherty, D., Kilmartin, B., Bourke, B., 2001. *Campylobacter upsaliensis* exerts a cytolethal distending toxin effect on HeLa cells and T lymphocytes. *Microbiology* 147, 735–743.
- Ogata, H., et al., 2001. Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. *Science* 293, 2093–2098.
- On, S.L., 2001. Taxonomy of *Campylobacter*, *Arcobacter*, *Helicobacter* and related bacteria: current status, future prospects and immediate concerns. *Symp. Ser. Soc. Appl. Microbiol.* 15–155.
- On, S.L., Harrington, C.S., 2001. Evaluation of numerical analysis of PFGE-DNA profiles for differentiating *Campylobacter fetus* subspecies by comparison with phenotypic, PCR and 16S rDNA sequencing methods. *J. Appl. Microbiol.* 90, 285–293.
- Pacheco, L.G., et al., 2011. A combined approach for comparative exoproteome analysis of *Corynebacterium pseudotuberculosis*. *BMC Microbiol.* 11, 12.
- Raskin, D.M., Seshadri, R., Pukatzki, S.U., Mekalanos, J.J., 2006. Bacterial genomics and pathogen evolution. *Cell* 124, 703–714.
- Rathinavelu, S., Kao, J.Y., Zavros, Y., Merchant, J.L., 2005. *Helicobacter pylori* outer membrane protein 18 (Hp1125) induces dendritic cell maturation and function. *Helicobacter* 10, 424–432.
- Rodriguez-Ortega, M.J., et al., 2006. Characterization and identification of vaccine candidate proteins through analysis of the group A *Streptococcus* surface proteome. *Nat. Biotechnol.* 24, 191–197.
- Ruiz, J.C., et al., 2011. Evidence for reductive genome evolution and lateral acquisition of virulence functions in two *Corynebacterium pseudotuberculosis* strains. *PLoS One* 6, e18551.
- Samant, S., et al., 2008. Nucleotide biosynthesis is critical for growth of bacteria in human blood. *PLoS Pathog.* 4, e37.
- Schlichtman, D., Kubo, M., Shankar, S., Chakrabarty, A.M., 1995. Regulation of nucleoside diphosphate kinase and secreted virulence factors in *Pseudomonas aeruginosa*: roles of algR2 and algH. *J. Bacteriol.* 177, 2469–2474.
- Shoaf-Sweeney, K.D., Larson, C.L., Tang, X., Konkel, M.E., 2008. Identification of *Campylobacter jejuni* proteins recognized by maternal antibodies of chickens. *Appl. Environ. Microbiol.* 74, 6867–6875.
- Skirrow, M.B., 1994. Diseases due to *Campylobacter*, *Helicobacter* and related bacteria. *J. Comp. Pathol.* 111, 113–149.
- Snipen, L., Ussery, D.W., 2010. Standard operating procedure for computing pangene trees. *Stand. Genomic Sci.* 2, 135–141.
- Soares, S.C., et al., 2012. PIPS: pathogenicity island prediction software. *PLoS One* 7, e30848.
- Spence, R.P., et al., 2011. Cross-reaction of a *Campylobacter fetus* subspecies *venerealis* real-time PCR. *Vet. Rec.* 168, 131.
- Stynen, A.P., et al., 2011. Complete genome sequence of type strain *Campylobacter fetus* subsp. *venerealis* NCTC 10354T. *J. Bacteriol.* 193, 5871–5872.
- Sun, J., Wang, X., Lau, A., Liao, T.Y., Bucci, C., Hmama, Z., 2010. Mycobacterial nucleoside diphosphate kinase blocks phagosome maturation in murine RAW 264.7 macrophages. *PLoS One* 5, e8769.
- Takamiya, M., et al., 2011. Genome sequence of *Campylobacter jejuni* strain 327, a strain isolated from a turkey slaughterhouse. *Stand. Genomic Sci.* 4, 113–122.
- Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28, 2731–2739.
- Tettelin, H., et al., 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome”. *Proc. Natl. Acad. Sci. U. S. A.* 102, 13950–13955.
- Thompson, S.A., 2002. *Campylobacter* surface-layers (S-layers) and immune evasion. *Ann. Periodontol.* 7, 43–53.
- Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680.
- Thompson, S.A., Shedd, O.L., Ray, K.C., Beins, M.H., Jorgensen, J.P., Blaser, M.J., 1998. *Campylobacter fetus* surface layer proteins are transported by a type I secretion system. *J. Bacteriol.* 180, 6450–6458.
- Tu, Z.C., Wassenaar, T.M., Thompson, S.A., Blaser, M.J., 2003. Structure and genotypic plasticity of the *Campylobacter fetus* sap locus. *Mol. Microbiol.* 48, 685–698.
- Wallden, K., Rivera-Calzada, A., Waksman, G., 2010. Type IV secretion systems: versatility and diversity in function. *Cell. Microbiol.* 12, 1203–1212.
- Whitehouse, C.A., Balbo, P.B., Pesci, E.C., Cottle, D.L., Mirabito, P.M., Pickett, C.L., 1998. *Campylobacter jejuni* cytolethal distending toxin causes a G2-phase cell cycle block. *Infect. Immun.* 66, 1934–1940.
- Wyszynska, A., Raczkowski, A., Lis, M., Jagusztyń-Krynica, E.K., 2004. Oral immunization of chickens with avirulent *Salmonella* vaccine strain carrying *C. jejuni* 72Dz/92 cjaA gene elicits specific humoral immune response associated with protection against challenge with wild-type *Campylobacter*. *Vaccine* 22, 1379–1389.
- Yamamoto, M., et al., 2011. Identification of genes involved in the glycosylation of modified viosamine of flagellins in *Pseudomonas syringae* by mass spectrometry. *Genes* 2, 16.
- Zhou, C.E., Smith, J., Lam, M., Zemla, A., Dyer, M.D., Slezak, T., 2007. MvirDB—a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic Acids Res.* 35 (Suppl. 1), D391–D394.

2.2 Conclusion

We were able to estimate the pangenome of the genus *Campylobacter*. This pangenome is considered to be high due to greater genomic variability and diverse species in the genus. Phylogenetic analysis revealed that *C. fetus* subspecies are distant from *C. jejuni* species. However, they were located in center between *C. jejuni* species and non-*jejuni* species. Considering *C. fetus* subspecies, it is found that they show 76% similarity in proteomic content. However, 428 and 88 strain specific gene families were recorded in *C. fetus venerealis* and *fetus* respectively. The core genome for both species contains 1,628 gene families. Functional characterization of the core genome and detail analysis of the sequences resulted in the identification of various potential virulence factors and vaccine candidates such as: nucleoside diphosphate kinase (Ndk), type IV secretion systems (T4SS), outer membrane proteins (OMP), surface array proteins, and cytolethal distending toxin (CDT). Unique genomic regions (pathogenicity islands) associated with *C. fetus* subspecies were shown to harbor: CRISPR loci and *dam* genes in an island specific for *C. fetus* subsp. *fetus*, and T4SS and *sap* genes in an island specific for *C. fetus* subsp. *venerealis*, respectively. The common and unique virulence factors and vaccine candidate data can be incorporated to public pathogenomic databases. However, we suggest more *Campylobacter fetus* species genomes to be sequenced for detailed knowledge of species pangenome, virulence factors and understanding the life style and detailed mechanism of pathogenicity.

CHAPTER 3
HELICOBACTER
(H. pylori)

3.1 Research Article

Computational comparative genomic based insights into human gastric pathogen *Helicobacter pylori* (38 species), essential features and species pangenome. (Manuscript).

In the following article, we explore the genomic information of an important human pathogen *Helicobacter pylori* which is the major cause of peptic ulcer and is recognized as the second leading cause of gastric cancer, in about half of the human population. Due to its immediate importance and significant concerns in health sciences, we attempted to determine the conserved genomic regions of the genus and species diversity. All the available complete genome sequences of *H. pylori* species (38) were extracted from public databases. Protein coding genes and rRNA genes were predicted, analysed and compared for genomic variations. The phylogenomic analysis, whole genome/proteome alignments, conserved core and pangenome were determined. The core genome of the species is then classified to functional categories based on homology searches with Cluster of Orthologous Genes (COG). Essential gene families were predicted by homology based searches against the database of essential genes and analysed for their homologs in host (human) proteome. Finally, the species specific virulence factors and candidate sequences were identified for therapeutic purposes and vaccine development.

Computational comparative genomic based insights into human gastric pathogen *H. pylori* (38 species), essential features and species pangenome.

Amjad Ali¹, Siomar C Soares¹, Syeda M Bakhtiar^{1,5}, Sandeep Tiwari¹, Syed S. Hassan¹, Fazal Hanan⁶, David W Ussery², Antaripa Bhattacharya⁴, Debmalya Barh⁴, Artur Silva³, Anderson Miyoshi¹, Vasco Azevedo^{1*}

¹ Laboratory of Cellular and Molecular Genetics, Federal University of Minas Gerais, Belo Horizonte, 31907-270, Minas Gerais, Brazil.

² Center for Biological sequence Analysis, CBS, Technical University of Denmark.

³ Federal University of Pará, Belém, 66075-110, Pará, Brazil.

⁴ Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur, WB-721172, India.

⁵ Mohammad Ali Jinnah University, Sehala Road, Islamabad.

⁶ KIMS, Khyber Medical University, Peshawar, Pakistan

***Corresponding Author:**

Dr. Vasco Azevedo. vasco@icb.ufmg.br

Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais. Av. Antônio Carlos 6627, Pampulha 31.270-901, Belo Horizonte, Minas Gerais, Brazil. Phone Number: + 00 55 31 3409-2610 and Fax Number: +00 55 31 3409-2610. URL Homepage: <http://www.icb.ufmg.br>

Abstract

H. pylori is the major cause of peptic ulcer in human and it is recognized as the second leading cause of gastric cancer (~70%), in about half of the human population. Conversely, an increased resistance to antibiotics and slow development in vaccines against *H. pylori* are observed. The comparative analysis of available complete genomic sequences of global representative *H. pylori* species on public databases consisting of 38 *H. pylori* (plus 8 non-*pylori*) species are presented here. The genomic variations/statistics, phylogenomic analysis, whole genome/proteome alignments, conserved core, and pan genome were determined. We characterized 1,185 conserved core gene families (CGFs), which represent ~77% of the average genome size and 42% of the species pan gene families (PGFs). The core genome of genus, *pylori* and non-*pylori* species, were then classified to biological and functional categories based on COG super-functional categories. Furthermore, 250 core essential gene families (EGFs) were predicted and explored for the presence of their homologs in host genome and potential candidate sequences were extracted. Beside this, 3 highly conserved pathogenicity islands and 2 highly variable islands have been predicted in *H. pylori* strains. The genomic variations, common features, and core EGFs could provide insight into the identification of species unique/shared pathways, virulence factors and to more efficient use of the candidates for vaccine, antibiotic, and drug development.

Key words

H. pylori, comparative genomics, core genome, pangenome, essential genes.

Abbreviations

GF, gene families; CGFs, core gene families; PGFs, pan genes families; EGFs, essential gene families.

Background

The genus *H.* contains bacterial species which colonizes the gastrointestinal tract of human and other mammals both *H. pylori* and non-*pylori* species are involved in gastrointestinal diseases (Lehours et al., 2011; You et al., 2012). Particularly in human, *H. pylori* causes diseases like gastritis and peptic ulcers, which can lead to the development of gastric cancer (~10% cancer deaths). It was therefore enlisted as a class I carcinogen by WHO, In 1994 (Dong, 2009; You et al., 2012). The prevalence of the disease remains high in developing world and poor socio-economic countries like the ones located in Africa, East Asia and Central America, where the diseases peptic ulcer and gastric cancer are highly prevalent (S. Zhang, Moise, & Moss, 2011). However, in the developed world and industrialized cities the prevalence is considerably lower, most probably due to preventive and public health measures in some western countries. Previous reports indicate that most of the *H. pylori* isolates could be classified either by sequence diversity or gene content. It has also been observed that *H. pylori* differs between two individuals and certainly greater differences could be seen in isolates from different continents. These genomic variations are indicative of genetic drift during geographic isolation and adaptation and co-evolution of the pathogen with different ethnic groups of humans (Gressmann et al., 2005; Mikkonen, 2004). To explore the common features and shared genomic contents, Salama et al., in 2000, determined the core set of genes in 15 *H. pylori* strains by microarray method and observed that 1281 genes are commonly shared by all examined strains. However, the examined strains were mainly isolated from Western countries (N. R. Salama et al., 2007; N. Salama et al., 2000). For detailed view of the diversity and conservation in *H. pylori* strains, Gressmann et al., in 2005, determined the core genome for a larger and global representative group of strains and suggested that *H. pylori* core genome consists of 1,111 genes. Beside the genetic repertoire of the organism (*H. pylori*), the core minimal set of genes would be of interest too, which is required for maintaining the basic cellular life, indispensable for survival. (R. Zhang & Lin, 2009). From the therapeutic point of view, the identification and characterization of essential genes and minimal gene set for *H. pylori* species is an interesting strategy, both theoretically and experimentally, to understand the core requirements for cellular life. A practical example of the importance of minimal gene set characterization is in drugs development which usually designed against essential cellular processes and essential gene products of microbial cells as they are promising new targets for such drugs (Marcio et al., 2009). Beside essential genes and proteins (House-keeping proteins) required by pathogens for maintaining their crucial cellular life functions, the

class of virulence factors is also important. In the course of pathogenesis and development of disease these virulence factors (VFs) or molecules are produced by pathogens which enhance the capabilities of the pathogen to cause disease (Zheng et al., 2012). These virulence factors can be divided into seven groups: 1) Adhesins, facilitates microbial attachment to host, 2) Colonization, ability of the pathogen to colonize the host cells, in the case of *H. pylori*, colonize the gastric regions, 3) resistance factors or effectors that provide resistance to host defense mechanism, 4) Invasion, helped by factors that disrupt the membrane and facilitates process like endocytosis, 5) Toxins are factors which are produced by several bacteria which are harmful poisons to host cell and causes damages, 6) Polysaccharides (capsular) helps the pathogen in the protection from host responses, 7) Siderophores for iron uptake. Recently, we witnessed a number of global representative complete genome sequences of *H. pylori* on genbank. Hence, with all those available complete genome sequences, we aimed to get insights into the genomic variation in *H. pylori* species, and to compare the genomes and predicted proteomes for estimating the conserved regions and to characterize them, particularly the essential features. Furthermore, we also aimed to predict the potential core immunogenic virulence factors and putative vaccine candidates against *H. pylori*.

Material and Methods

Data Collection and Management

The global gastric pathogen *H. pylori* have multiple complete genomes available on public databases for scientific exploitation. We, therefore, selected a total of 38 complete *H. pylori* genomes and 8 non-*pylori* *H.* species. Genbank (gbk) files were obtained from NCBI genome browser (<http://www.ncbi.nlm.nih.gov/genome/browse/>). As a starting point, DNA sequences in fasta format were extracted from all the genbank files (chromosome and plasmid sequences were gathered, as they were required) and then subjected to program **RNAmmmer** (Lagesen et al., 2007) for prediction of full length 16S rRNA gene sequences. The program **Prodigal**, a gene finding algorithm, was also used to predict Open Reading Frames (ORFs) in all genomes (Hyatt, LoCascio, Hauser, & Uberbacher, 2012). We, however, have seen slight variations in the number of genes/ORFs in our predicted ones and those of public predicted (genbank). To avoid inconsistencies and to get uniformity in the resultant data, we used single gene finding program. **Table 1** demonstrates the basic genome information i.e. size in base pairs, number of predicted proteins, chromosome/plasmid accession numbers and percent AT content.

Phylogenomic Analysis

As a starting point, for better understanding the evolutionary relationships and genomic variations at genus level (*H. pylori* and non-*pylori* species), the program **Gegenees** (version 1.1.4) was used, and the genbank files were retrieved from genbank for all the *H.* species (Table 1)(Agren, Sundström, Håfström, & Segerman, 2012). The program splits the genomic sequences into smaller parts and similarity searches are performed by all-vs-all alignments. The variable contents generated are then compared with all other genomes. Finally, the percentages of similarity are recorded and a heatmap is generated. For visual inspection, color intensity in the heatmap present the degrees of similarity. Beside the heatmap chart, a nexus file was extracted from Gegenees for input into software **SplitsTree** (4.12.6) with an option of generating phylogenetic tree by Neighbor-Joining Method (NJ)(Huson, 1998). Additionally, a 16S rRNA gene sequences based tree was also constructed and is shown in **Additional Figure 1**.

Proteome Comparisons (Pairwise alignment)

All the selected *H. pylori* genomes were translated to their proteomes. BLASTp comparisons were carried out for all the proteins in one genome against all the proteins in the other genomes in the study (Ali et al., 2012; Lukjancenko, Ussery, & Wassenaar, 2012). The blast parameters were set as follows: blast e-value of 1e-5; homologs cutoff of 1e-8; length of homolog of 0.95; identity homolog of 0.95; the blast minimum score of 30; and, minimum identity of 90. Reciprocal best blast hits were selected for generation of blast matrix which is shown in **Figure 2**. Each cell in the blast matrix shows how many genes/proteins (X-axis) had reciprocal hits with respect to the genome listed to the right (Y-axis). Each corresponding rectangular box (between any two genomes) in the matrix represents the number of reciprocal hits (shared proteins) and the total number of proteins. For example, two genomes, G1 and G2, will lead to $(G1/G2)*100 = X\%$ of the proteins/genes in G1 had reciprocal hits, while $(G2/G1)*100 = X\%$ of the proteins/genes in G2 had reciprocal hits. The diagonal row of boxes indicates the internal homologies against itself (genome). The colored matrix is generated for all the 38 genomes/proteomes with the scale given, indicating the relative homology between corresponding genome/proteome.

Pangenome Analysis and Functional Characterization of Core genome.

The *H. pylori* conserved core families (CGFs) and pangenome families (PGFs) were estimated followed by previously established method (Ali et al., 2012; Lukjancenko et al., 2012). CGFs and PGFs were estimated by employing single-linkage clustering on

top of BLASTp alignments, with the notion that any two genes in the data set are considered to belong to the same gene family and should be considered as 'conserved' once their amino acid sequence is at least 50% identical over at least 50% of the length of the longest gene (Leekitcharoenphon, Lukjancenko, Friis, Aarestrup, & Ussery, 2012; Lukjancenko et al., 2012). Doing so, multiple genes may belong to single gene family and the number of gene families would be lower than the actual number of genes in a genome. Those genes which do not fit to the criterion would constitute individual genes families. Gene families with at least one gene in common were gathered into the core genome. The rest of the total, either unmatched or not qualifying according to the criterion, constitutes the pan genome. The core genes of genus *H.* as a whole, *H. pylori* species and non-*pylori* species, were categorized according to the Cluster of Orthologous Genes (COG) super-functional category in: 1. Information storage and processing; 2. Cellular processes and signaling; 3. Metabolism; and 4. Poorly characterized. To perform this analysis, protein similarity searches (blastp) were made for the core sets of genes against the COG database (<http://www.ncbi.nlm.nih.gov/COG/>). The proteins with an e-value higher than 10^{-6} were excluded, and the best blast results for each protein were considered for COG super-functional categories (Tatusov, Galperin, Natale, & Koonin, 2000).

Essential gene families and non-host homologs prediction

The *H. pylori* predicted core genome was then aligned with the Database for Essential Genes, DEG (<http://tubic.tju.edu.cn/deg/deg.rar>), for estimation of essential core gene families, EGFs (R. Zhang & Lin, 2009). DEG contains essential genes data from more than 10 bacteria, for example, *E. coli*, *B. subtilis*, *H. pylori*, *S. pneumoniae*, *M. genitalium* and *H. influenzae*. The BLAST comparison settings for selecting the essential genes/ proteins were followed as previously described (Butt, Nasrullah, Tahir, & Tong, 2012): expected value (E-value) cut-off of 10^{-10} ; percentage of identity $\geq 35\%$ between query and hits; and, minimum bit score of 100. In the case of *H. pylori*, the host (human) genome/proteome was downloaded from NCBI (*taxid*: 9606) and BLASTp analysis were performed. In BLASTp comparison parameters, the percentage of identity and E value were kept $<35\%$ and 0.005 respectively. Proteins without hits below the E-value inclusion threshold were collected as non-host bacterial proteins and we called them non-host essential gene families (nHEGFs).

Vaccine and drug targets Prediction

The prediction of potential immunogenic and vaccine candidate gene and proteins the predicted nHEGFs (250) were analysed for their surface localization. The pipeline

SLEP (Surface Localization Extracellular Proteins) was used for this purpose (Giombini, Orsini, Carrabino, & Tramontano, 2010). The pipeline integrates the results from various programs (Glimmer, TMHMM, PRODIV-TMHMM, LipoP, PSortB) (Giombini et al., 2010). Furthermore, for prediction of potential vaccine targets and analyses, a web tool/pipeline **Vaxign** (<http://www.violinet.org/vaxign/>) (He, Xiang, & Mobley, 2010) was used. The set of surface proteins were subjected to dynamic search tools with specific parameters: adhesin probability \geq 0.51, transmembrane helices \geq 1, MHC class I and II binding epitopes, non-host conservation and pathogen specific proteins. Vaxign software already have analysed (validation) data from total of 254 genomes (bacteria) and having 5,80,900 proteins. This also including data from six *H. pylori* genomes (*H. pylori* 26695, *H. pylori* G27, *H. pylori* HPAG1, *H. pylori* J99, *H. pylori* P12 and *H. pylori* Shi470) with total of 9,261 proteins already analysed and incorporated into vaxign database. For drug target identification the core non-host essential proteins (nHEGFs) were analysed by subcellular localization software **PSORTb** v3.0.2 (Gardy et al., 2005) with default settings. The obtained set of targets proteins were then structurally analysed using **Phyre** v2.0 (Protein homology/analogy recognition engine) (<http://www.sbg.bio.ic.ac.uk/phyre2/>) and NCBI BLASTp with PDB (Protein Data Bank) (<http://www.rcsb.org/pdb/home/home.do>), with threshold 1. Phyre uses a library of known protein sequences derived from the Structural Classification of Proteins (SCOP) database and from the Protein Data Bank (PDB). These sequences are used to construct a non-redundant fold library. Each sequence in the fold library is iteratively scanned against a non-redundant sequence database and a hidden Markov model (HMM) for each known structure generated. This fold library also contains known and predicted secondary structures for all protein sequences stored. Similarly, protein sequences submitted are scanned against the non-redundant sequence database, and a profile HMM is created. PSI- Blast is used to collect both close and remote sequence homologues, and an alignment is constructed. Following this, secondary structure is predicted. The profile HMM and the secondary structure are then used to scan the fold library using HMMeHMM matching. This alignment process returns a score on which all alignments are ranked, and an E-value is generated (Kelley & Sternberg, 2009).

Pathogenomic analysis and pathogenicity islands prediction

In order to identify pathogenicity islands (PAIs) on *H. pylori* species, we have performed analysis with PIPS: Pathogenicity Island Prediction Software (Soares et al., 2012). PIPS identify PAIs according to their main feature: genomic signature deviations, i.e., G+C content and codon usage deviations; presence of virulence

factors, transposase genes and flanking tRNAs genes; and, absence in non-pathogenic organism from the same genus or related species. The analysis were performed on *H. pylori* strains 26695, Cuz-20, J99, PeCan4 and SouthAfrica7, which are representative of Europe, East Asia, West Africa, South America and South Africa, respectively. Additionally, *Wolinella succinogenes* DSM 1740 was chosen as non-pathogenic species for PIPS analysis (Baar et al., 2003). After prediction step, the sizes of all PAIs in all 5 *H. pylori* genomes used in this step were curated through graphical genome comparisons in ACT: the Artemis Comparison tool (Carver et al., 2005). Finally, we have obtained reference PAIs from the genomes where they were more representative (accordingly to their size and gene content), compared the reference PAIs with all 38 genomes of *H. pylori* in this work using the proteome comparison data and plotted the percentages of similarity in a heatmap for easy of view.

Results and Discussion

H. pylori General Features and Genome Statistics

H. pylori (*H. pylori*), is Gram-negative micro-aerophilic bacterium colonizes the human stomach, a pathogens belonging to epsilon-bacteria. (Eppinger et al., 2006; Lehours et al., 2011; Tomb et al., 1997). The bacteria have a total of 38 complete (RefSeq) genomes on NCBI genbank (at the time of analysis). There are also 8 *H. non-pylori* species available. In 38 *H. pylori* complete genomes, about half of them (17) carry plasmid (**Table 1**). Out of 8 non-*pylori* species, 5 have plasmid. The total number of proteins in 38 *H. pylori* genomes was calculated as 58,379, while an average genome contains 1,536 genes/proteins. The first genome sequence of *H. pylori* 26695 was obtained in 1997, consisting of a circular chromosome with an average GC content of 39% and 1,667,867 base pairs (Tomb et al., 1997). According to genbank data, the %GC does not have considerable variations and range in 38-39% (<http://www.ncbi.nlm.nih.gov/genome/browse/>). In *H. pylori* species, an average 61% AT content was observed. The closely related strain J99 isolated from an American patient having duodenal ulcer, when compared to *H. pylori*, 26 695, total of 1406 genes were found common in them. The J99 was observed to have 86 unique genes, however, both strains share the *cag* pathogenicity island, which codes the type IV secretion system facilitating the transport of *CagA* cytotoxin to gastric epithelial cells (Dong, 2009; Kutter et al., 2008). Later, the strain HPAG1 (isolated from an elderly women in Sweden) sequenced in 2006 was compared to the previous two strains (J99 and 26 695) and it has smaller genome size and shared the *CagA* and *vacA* (virulent allele). HPAG1 was found to contain 43 strain specific genes. Another strain, G27 strain, which is similar in size (1652983 bp) to the previous strains

(26 695, J99 and HPAG1) contains a plasmid (11 genes) and 58 specific genes. However, the *cag* island is reported to be disrupted by a transposon in G27 (Amieva et al., 2003; Dong, 2009). The strain Shi470 have a genome relatively larger in size (~1.6 Mbp) than the previous strains. There is consensus that increase in genomic variability in *H. pylori* genomes and the phenomena of mutations and recombination are continuously occurring. For a large set of genomes (38 *H. pylori* plus 8 non-*pylori*) and their shared genomes and gene associated/specific to several strains are predicted and shown in **Table 1**. The highest number of strain specific genes (216 GFs) were observed in the *H. pylori* strain 26695 while the lowest (9 GFs) are in *H. pylori* Shi169.

***H. pylori* and non- *pylori* Phylogenomic Analysis**

H. pylori is a well recognized human pathogen, being colonizing the earliest human populations, and, it is supposed that they have co-evolved and co-migrated across the world (S. Zhang et al., 2011). A phylogenomic analysis has been done for 46 available *H.* species including 38 *H. pylori* and 8 non-*pylori*. The percentage of similarities calculated between species (Material and Methods) are demonstrated in heatmap (**Figure 1**). The non-*pylori* species shows greater genomic variation with *pylori* species and interestingly they do the same between each other (red color). On the other side, as expected, the *H. pylori* genomes showed a higher percentage of similarities between them, and the greenish blocks indicated the same followed by the yellow one. On the left side, the tree illustrates 2 defined clades, representing between *H. Pylori* species with greater similarity/relatedness in their genomic contents, while non-*pylori* species positioned bit far and gathered on separate branches depending on intra-genomic divergence. Among the non-*pylori* species, *H. acinonychis* shows greater genomic similarity with *H. pylori* species and was found to share 612 orthologues, suggesting that *H. acinonychis* has recently shifted from human to felines (Eppinger et al., 2006). Beside the heatmap analysis, we constructed a Tree based on 16S rRNA genes sequence divergence, which resulted in somewhat similar results (**Additional Figure 1**), where two clades with multiple branches can be seen, indicating the genomic relatedness and slight 16S rRNA sequence divergence among *H. pylori* isolates. The previous reports have shown that the *H.* species divides into two clusters: one with the species which colonize the stomach (so called gastric species) of mammals and the other with species that inhabit the intestine and biliary tracts (so called enterohepatic cluster) (Mikkonen, 2004).

Proteome conservation in *H. pylori* genomes and pairwise comparisons

As we observed higher similarities in genomic contents of *H. pylori* species in phylogenetic analysis, we decided to compare the whole genome/proteome of all *H. pylori* species. On the other side non-*pylori* inter-species showed greater variations in genomic content and hence we did not include them for proteome analyses. The *H. pylori* genomes were pairwise compared (BLASTp). To estimate the amount of shared proteome between any two species and to visualize the results a Matrix was generated from the data of comparisons which is shown in **Figure 2**. Both number/amount and percentage of shared proteome (proteins) is shown in corresponding box of the matrix. We can see in the color matrix, a minimum proteome conservation/homology of ~80% (80.59) and a maximum of ~98% (98.27). The maximum homologies we observed in between *H. pylori* 2017 and *H. pylori* 2018 and on the other way around, and *H. pylori* 908 and *H. pylori* 2017. On the other hand, one can see the minimum ~81% proteome homology between *H. pylori* 2018 and *H. pylori* Shi 470. Similarly, *H. pylori* 908 and *H. pylori* B8 sharing the same amount of proteins with *H. pylori* Shi 470. These findings are interesting when compared with phylogenetic trees and reflects the organisms relationships. For example, *H. pylori* 2017 and *H. pylori* 2018 are located quite near (phylogenomic analysis), this also reflects the geographic similarity. They are closely related to *H. pylori* 908, whereas *H. pylori* Shi 470 is located on the other clade and on one of the sub branch.

Pangenome analysis of *Helicobacter* (*pylori* and non-*pylori*)

The predicted core or central genome of *H. pylori* species contains 1,185 CGFs, and the pangenome contains 2,825 PGF. An average genome of *H. pylori* contains ~1536 genes (lowest 1,464 in *H. pylori* Sat464 and highest 1,701 in *H. pylori* XZ274), and the total number of genes predicted in 38 *Pylori* genomes was 58,379. The calculated core genome represents ~77% of the average genome size (1,536 genes/proteins). As expected, the *H. pylori* species share greater part of their genomic contents and a decreasing number of new GFs/clusters was observed to what was accumulated upon subsequent genome incorporation to the study. We observed, on average, ~46 new gene clusters/GFs accumulates with subsequent addition of each genome; however, the highest numbers of species specific GFs observed are 206, 116, 100 in *H. pylori* 26695, *H. pylori* 35A and *H. pylori* XZ274, respectively (**Table 1**). The pangenome, on the other side, raised at a lower rate, roughly, *H. pylori* pangenome is twice the size (58%) of the core genome. In a previous study by Salama et al., in 2000, they determined the core set of genes in 15 *H. pylori* strains with a microarray method, and total of 1,281 genes were found to constitute the core genome (N. Salama et al., 2000). As the strains were mainly isolated from only the western part of the

world, hence, it was difficult to estimate that those core genes would also represent the larger number of species, once they became available. However, it is not much surprising to estimate high number of CGFs for a few intra-species genomes due to greater genomic similarities. In a systematic comparison, the first 15 genomes in our dataset represent 1,220 CGFs, which is not that far from the core genome (1,281) observed by Salama et al.. Furthermore, 25 genomes shares 1,200 CGFs and the total set of 38 genomes constitutes a core genome of 1,185 CGFs (**Additional File 1**). On the other hand, we analysed 8 complete genome sequences (non-*pylori* genomes; Table 1) from the same genus and a much smaller core genome was observed, containing only 470 CGFs (**Additional File 2**) and the PGFs goes up to 6,472. Unlike *H. pylori*, there are few (8) non-*pylori* complete genomes available so far, and, even though, greater genomic variations are observed in inter-species comparisons. Also, an average of ~760 new GFs added to the study on addition of each new species. However, among the non-*pylori* species, *H. acinonych* shows higher similarity to *H. pylori* based on heatmap data. Also, previously reported that *H. acinonych* share 612 orthologues with only few amino acids differences. These findings also supports possible host shifts from human to felines (Eppinger et al., 2006). Beside the *pylori* and non-*pylori* species, we calculated the core genome of the genus as a whole (46 species), which contains 463 GFs, which is close to the core genome of the genus (470GFs). The *H. pylori* total number of gene clusters, core and pan genome bars pattern (ups and down) is shown in a chart in **Figure 3**. A pangenomic dendrogram has also been generated from the same data, based on presence or/and absence of specific GFs among *H. pylori* and non-*pylori* genomes and is shown in **Additional Figure 2**.

Functional categorization of core genome (genus, *pylori* and non-*pylori*)

Functional categorization of the core genomes: *H. pylori* species, non-*pylori* species, and as a whole (genus) were performed using COG super-functional categories. And, according to **Figure 4** (chart), the majority of the GFs/proteins (364, 173 and 172 respectively) are associated under the category “Metabolism”. In brief, GFs in this category performs important functions like, production and conversion of energy and transports of biomolecule, where metabolism and transport of molecules such as, nucleotides, amino acids, carbohydrates, coenzymes etc. are among the vital functions. Due to the high importance of the genes under the category metabolism, they are anticipated to be highly represented in core genome. The second major category in all the three set of data (core) is the “Cellular Process and Signaling”, members of this this category are associated with functions like: cell cycle control, cell

division, chromosome partitioning, cell wall/membrane/envelope biogenesis, cell motility, post-transcriptional modification, signal transductions, intracellular trafficking, secretion, defense mechanisms etc. The third super functional category, which is almost the same in number of GFs, is the “Information Storage and Processing”, where the contents of this category are involved in the process of translation, ribosomal structure and biogenesis, RNA processing and modification, DNA transcription and replication etc. While the last/4th category is “Poorly Characterized”, which contains GFs usually involved in general cellular functions or with undefined functions. Overall, the distribution of core genomes (CGFs) of *H. pylori*, non-*pylori* and genus *H.* , to similar categories are consistent (Tatusov et al., 2000).

***H. pylori* core essential gene families and host homologs predictions**

Essential genes are those genes which are crucial for growth, cellular activities and foundation of life. These genes represent a minimal gene set which is essential for a living cell. The identification of essential genes are of practical importance in drug designing against bacterial infections, and due to the fact that most of the antibiotics target essential cellular processes and essential gene products of microbial cells, they are candidate targets for such drugs (Acencio & Lemke, 2009). For this propose, we subjected our predicted CGFs (1,185) against the database of essential genes, DEG, and classified genes/CGFs as EGFs if they had significant homology against experimentally validated essential genes in other bacteria, bacterial group or database. We predicted a total of 493 (~42%) EGFs in our core data set with the following selection parameters: E-value cut-off 10^{-10} , percentage of identity $\geq 35\%$ between query and hits, and a minimum bit score of 100 (Butt et al., 2012). The list of those 493 EGFs along with their corresponding/counterparts with DEG ID are listed in **Additional File 3**. In principle, ideal targets for bacterial therapeutics are pathogen's genes/protein whose homologs are absent in corresponding host. In the case of *H. pylori*, colonizing human, we downloaded the human proteome (*taxid*: 9606), which has 34,521 proteins. BLASTp analysis were performed and the predicted core EGFs in previous step (493) were aligned against human proteome. This time, the E-value cutoff was set to 0.005 and the percent of identity to 35%. Proteins without hits below the set parameters were chosen as non-host bacterial proteins. Two datasets were then generated and the so-called pathogen essential gene families having human homologs (243) was kept aside. The remaining 250 Core essential proteins were found specific to pathogen (*H. pylori*), and do not have corresponding human homologs with same threshold (**Additional File 5**). The later set (core essential non-human homologous proteins = 250) were selected as potential candidates and subjected to further validations.

Vaccine candidates identification

For multiple pathogenic species (intra species), once the complete genome sequences are available, it is a desirable task to predict global vaccine targets that are conserved among all genomes. Generally, surface and secreted proteins are involved in bacterial pathogenesis and are recognized as good candidates for vaccine development (Ali et al., 2012; Giombini et al., 2010). On the other hand, cytoplasmic and inner membrane proteins may not be good candidates for vaccine development, however, they can be targets for drug designing against bacteria. For *H. pylori* global vaccine identification propose, we analysed our non-human homologous set of EGFs (250) predicted by **SLEP** pipeline (Giombini et al., 2010). 88 surface protein were predicted out of 250 EGFs (69 membrane, 12 exported and 7 lipoproteins) and the multi fasta sequences file, which contains 88 surface proteins was created and is given as **Additional File 5**. Furthermore, reverse vaccinology strategy and web-based software “Vaxign” for vaccine designing has been consulted. The program already have analysed data for 254 genomes and 5,80,900 proteins. The data from 6 *H. pylori* genomes (*H. pylori* 26695, *H. pylori* G27, *H. pylori* HPAG1, *H. pylori* J99, *H. pylori* P12 and *H. pylori* Shi470) and 9,261 proteins have also been analysed. The predicted 88 *H. pylori* (CGFs/EGFs) surface protein predicted by using **SLEP** were subjected to **Vaxign** dynamic search tool with settings (adhesin probability \geq 0.51, transmembrane helices \geq 1) (He et al., 2010). The resulted candidate sequences (proteins) were filtered and analysed by program **WebMGA** for functional annotations and prediction of their roles in biological processes (Wu, Zhu, Fu, Niu, & Li, 2011). The candidate proteins were found to classified to functional biological categories and having role in cellular activities (except one hypothetical). According to COGs and KOGs classification, the candidate proteins found have significant roles in essential biological processes: Intracellular trafficking and secretion (COG0848, COG1862 and COG3736). Cell wall/membrane/envelop biogenesis (COG0744, COG1452 and COG1519). Lipid metabolism (COG0671). Energy production and conversion (COG0711). Amino acid transport and metabolism (COG0757). Post-translational modification, protein turnover, chaperones (COG4736). The predicted COGs, annotation and functional class descriptions are given in **Table 2** and the protein sequences (Multifasta) for these candidates are also provided in **Additional File 6**.

Drug targets identification (structural analysis)

A set of 139 core-target proteins sorted by PSORTb were then analysed for fold recognition using Phyre2 web server (Kelley & Sternberg, 2009). The results obtained

from the Phyre2 were analyzed using the cutoff value: confidence score $\geq 75\%$ and query coverage $\geq 15\%$ to find out any fold matching with host (Mao et al., 2013). No significant hits were found with host proteome, however, maximum number of folds were observed with bacterial species. Among the organisms where fold were matched include *Mycobacterium*, *Campylobacteria*, *E.coli*, *Yersinia pestis*, *Burkholdenria*, *Neisseria* and with some parasite such as *Entamobea histolitica*. This analysis also validated that targets proteins are found in many pathogen and were not observed in normal microflora. Due to the fact that most antibiotics target pathogens as well as healthy bacteria in human microbiota, this leading to a prolonged impact on normal gut flora. Therefore it is important that the designed novel drugs should specifically target only pathogenic bacteria. To crosscheck the result we used NCBI BLASTp with PDB and after analysing the result with the same query coverage (15%), we found that 8 proteins are giving some hits with chain of human proteins so these targets were discarded. The final set of target proteins (131) were then functionally annotated and provided in **Additional file 7**.

Predicted pathogenicity islands in *H. pylori*.

We have identified 22 non-redundant PAIs across the genomes of *H. pylori* strains 26695, Cuz-20, J99, PeCan4 and SouthAfrica7, named putative pathogenicity islands of *H. pylori* 1-22 (PiHp1-22). To create a heatmap with the percentage of similarity of each PAI on all genomes, the content of PAIs identified as PiHp1-18, PiHp19, PiHp20-21 and PiHp22 were acquired from the genome sequences of *H. pylori* strains 26695, Cuz-20, SouthAfrica7 and PeCan4, respectively. Although some of the reference PAIs may be partially or totally present in the other reference genomes, these reference PAIs were chosen due to the higher representability (according to their size and gene content), e.g., PiHp19 of *H. pylori* Cuz20 is also present in *H. pylori* 26695, however, it has only 40% of the genes of PiHp19 of *H. pylori* Cuz20 and, thus, the later one has been chosen for comparative analysis.

From the heatmap, we observed that, a high degree of variability in most of the PAIs across all 38 genomes, where only PiHps 2 (C694_01095-C694_01190), PiHps 4 (C694_01445- C694_01490), PiHps 14 (C694_05735-C694_05795) and PiHps 15 (C694_06365-C694_06390) are totally present in at least 50% of the strains (**Figure 5**). On the other hand, PiHps 8 (C694_02175-C694_02345) and PiHps 13 (C694_05045-C694_05230) are the most variable regions, presenting percentages of similarity ranging from 0-57% and 10-71%, respectively.

The majority of the genes harbored by PiHps: 2, 4, 8, 13, 14 and 15 have been assigned the term “hypothetical protein” on their product tag, meaning that most of the

gene products are not yet identified. However, some genes on these PAIs deserve attention. PiHp2, harbors a gene coding for a DNA repair protein, RadA (C694_01125), which is part of a superfamily of recombinases or DNA strand-exchange proteins composed of archaeal RadA, bacterial RecA and eukaryal Rad51 and DMC1 proteins. RadA has a pivotal role in DNA strand-exchange process between single stranded DNA (ssDNA) and a homologous double-stranded DNA (ds-DNA) in homologous recombination (Du & Luo, 2012). **Figure 6** demonstrates the conserved PiHps: 2, 4, 14 and 15 (A,B,C and D) and variable PAIs PiHps: 8, 13 and 9 (E, F and G).

Homologous recombination is one of several DNA repair pathways (direct reversal, base excision repair, nucleotide excision repair, mismatch repair, and recombination repair pathways) and functions in the repair of double-stranded DNA breaks and the restarting of stalled replication forks, therefore, warranting accurate functioning and propagation of genetic information (Du & Luo, 2012; Morita et al., 2010). However, as RadA is mainly found in archaea, in vitro experiments are required to assay the coding and to elucidate the putative function of RadA in bacteria, specially *H. pylori*; and, to clarify its putative acquirement through horizontal gene transfer from archaea.

In PiHp4, one can find genes coding for a cell division protein FtsH (C694_01445) and a toxin-like outer membrane protein (C694_01460), also termed putative vacuolating cytotoxin (VacA)-like protein. The cell division protein FtsH is a member of the AAA+ super-family (ATPases associated with diverse cellular activities), which consists of highly conserved molecular machines responsible for a number of cellular processes like, cell division, cell differentiation, signal transduction, stress response and others (Moldavski, Levin-Kravets, Ziv, Adam, & Prag, 2012). FtsH is required for the proper functioning of sigma 54 under nitrogen limitation conditions and FtsH mediated degradation of misassembled membrane and cytoplasmic proteins is thought to be responsible for its role in the heat shock response, therefore, explaining its upregulation in response to heat shock in several bacteria, including *H. pylori* (Beier, Spohn, Rappuoli, & Scarlato, 1997; Kiran et al., 2009)

VacA, or Vacuolating cytotoxin A, is one of the most studied toxins produced by *H. pylori*, whereas the CagA (cytotoxin-associated gene A) occupies the first place. VacA has the ability to cause vacuole-like membrane vesicles in the cytoplasm of gastric cells and is also associated with disruption of mitochondrial functions, stimulation of apoptosis and blockade of T-cell proliferation. The presence of the toxigenic allelic s1 form of VacA in strains of *H. pylori* is commonly associated with an increased risk of peptic ulceration and gastric cancer (Cover, Hanson, & Heuser, 1997; Jones, Whitmire, & Merrell, 2010; Kuck et al., 2001; Palframan, Kwok, & Gabriel, 2012). CagA, on the other hand, after injected in host cells, can influence: cellular tight junction, cellular

polarity, cell proliferation and differentiation, cell scattering, and induction of the inflammatory response. CagA was already identified to be located in an extensively studied pathogenicity island (cag-PAI), which is ~40 kb in size, encodes for a type IV secretion system (T4SS) and was identified by PIPS, in this work, as PiHp9 (C694_02670-C694_02870) (for detailed reviews of VacA and CagA, please refer to (Palframan et al., 2012) and (Cover et al., 1997).

Inside the most variable PAIs, PiHps 8 and 13, IS605 transposases tnpA (C694_02220, C694_05100 and C694_05150) and tnpB (C694_02225, C694_05105 and C694_05145) can be observed, which were probably responsible for the incorporation of the PAIs and could also account for their high instability. Additionally, PiHp8 also harbors genes coding for 2 VirB4-like proteins (C694_02240 and C694_02345). VirB4 is part of T4SS in bacteria, which is necessary for pilus biogenesis, substrate transfer, and virulence as it is responsible for horizontal transfer of plasmid DNA between bacteria during conjugation and for the delivering of macromolecules to prokaryotes and eukaryotes (Mossey, Hudacek, & Das, 2010; Walldén et al., 2012). In *H. pylori*, the T4SS formed by the VirD4/VirB genes are known to have a pivotal role in the delivery of CagA protein into the cytosol of gastric epithelial cells and the entire T4SS is coded by VirD4/VirB genes harbored by the cag-PAI (PiHp9). The putative backup function of the 2 copies of VirB4-like proteins in PiHp8 for the VirB4 harbored by cag-PAI is yet to be elucidated.

Conclusion

The study illustrates the comparative genomic and pangenomic analysis of an important human pathogen *H. pylori*. Multiple genome from the same species have been analysed, shown to contain greater intra-species genomic similarities and constitutes relatively higher conserved core genome. The number of strain specific genes were observed to be considerable low, indicative of the close relationships among the species. We provide insight into the pathogenesis of gastric pathogen *H. pylori*, a number of vaccine candidates and drug targets have been characterized. Furthermore, conserved pathogenic regions in multiple genomes have been identified which will helps in understanding the common pathogenic behavior of the pathogen. Finally, the comparative genomic tools and technique applied in the study can be extended to other pathogens even on larger scale. We believe that the data generated during this study can assist further research particularly in diagnosis, antibiotics, and vaccine development against *H. pylori*.

References;

- Acencio, M. L., & Lemke, N. (2009). Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. *BMC bioinformatics*, 10, 290. doi:10.1186/1471-2105-10-290
- Agren, J., Sundström, A., Håfström, T., & Segerman, B. (2012). Gegenees: fragmented alignment of multiple genomes for determining phylogenomic distances and genetic signatures unique for specified target groups. *PloS one*, 7(6), e39107. doi:10.1371/journal.pone.0039107
- Ali, A., Soares, S. C., Santos, A. R., Guimarães, L. C., Barbosa, E., Almeida, S. S., Abreu, V. A. C., et al. (2012). *Campylobacter fetus* subspecies: Comparative genomics and prediction of potential virulence targets. *Gene*. doi:10.1016/j.gene.2012.07.070
- Amieva, M. R., Vogelmann, R., Covacci, A., Tompkins, L. S., Nelson, W. J., & Falkow, S. (2003). Disruption of the epithelial apical-junctional complex by *Helicobacter pylori* CagA. *Science (New York, N.Y.)*, 300(5624), 1430–4. doi:10.1126/science.1081919
- Baar, C., Eppinger, M., Raddatz, G., Simon, J., Lanz, C., Klimmek, O., Nandakumar, R., et al. (2003). Complete genome sequence and analysis of *Wolinella succinogenes*. *Proceedings of the National Academy of Sciences of the United States of America*, 100(20), 11690–5. doi:10.1073/pnas.1932838100
- Beier, D., Spohn, G., Rappuoli, R., & Scarlato, V. (1997). Identification and characterization of an operon of *Helicobacter pylori* that is involved in motility and stress adaptation. *J. Bacteriol.*, 179(15), 4676–4683. Retrieved from <http://jb.asm.org/content/179/15/4676>
- Butt, A. M., Nasrullah, I., Tahir, S., & Tong, Y. (2012). Comparative Genomics Analysis of *Mycobacterium ulcerans* for the Identification of Putative Essential Genes and Therapeutic Candidates. (C. K. Carlow, Ed.) *PLoS ONE*, 7(8), e43080. doi:10.1371/journal.pone.0043080
- Carver, T. J., Rutherford, K. M., Berriman, M., Rajandream, M.-A., Barrell, B. G., & Parkhill, J. (2005). ACT: the Artemis Comparison Tool. *Bioinformatics (Oxford, England)*, 21(16), 3422–3. doi:10.1093/bioinformatics/bti553
- Cover, T. L., Hanson, P. I., & Heuser, J. E. (1997). Acid-induced dissociation of VacA, the *Helicobacter pylori* vacuolating cytotoxin, reveals its pattern of assembly. *The Journal of cell biology*, 138(4), 759–69. Retrieved from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2138037&tool=pmcentrez&rendertype=abstract>
- Dong, Q.-J. (2009). Comparative genomics of *Helicobacter pylori*. *World Journal of Gastroenterology*, 15(32), 3984. doi:10.3748/wjg.15.3984
- Du, L., & Luo, Y. (2012). Structure of a hexameric form of RadA recombinase from *Methanococcus voltae*. *Acta crystallographica. Section F, Structural biology and crystallization communications*, 68(Pt 5), 511–6. doi:10.1107/S1744309112010226

- Eppinger, M., Baar, C., Linz, B., Raddatz, G., Lanz, C., Keller, H., Morelli, G., et al. (2006). Who ate whom? Adaptive *Helicobacter* genomic changes that accompanied a host jump from early humans to large felines. *PLoS genetics*, 2(7), e120. doi:10.1371/journal.pgen.0020120.eor
- Gardy, J. L., Laird, M. R., Chen, F., Rey, S., Walsh, C. J., Ester, M., & Brinkman, F. S. L. (2005). PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics (Oxford, England)*, 21(5), 617–23. doi:10.1093/bioinformatics/bti057
- Giombini, E., Orsini, M., Carrabino, D., & Tramontano, A. (2010). An automatic method for identifying surface proteins in bacteria: SLEP. *BMC bioinformatics*, 11, 39. doi:10.1186/1471-2105-11-39
- Gressmann, H., Linz, B., Ghai, R., Pleissner, K.-P., Schlapbach, R., Yamaoka, Y., Kraft, C., et al. (2005). Gain and loss of multiple genes during the evolution of *Helicobacter pylori*. *PLoS genetics*, 1(4), e43. doi:10.1371/journal.pgen.0010043
- He, Y., Xiang, Z., & Mobley, H. L. T. (2010). Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development. *Journal of biomedicine & biotechnology*, 2010, 297505. doi:10.1155/2010/297505
- Huson, D. H. (1998). SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics (Oxford, England)*, 14(1), 68–73. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/9520503>
- Hyatt, D., LoCascio, P. F., Hauser, L. J., & Uberbacher, E. C. (2012). Gene and Translation Initiation Site Prediction in Metagenomic Sequences. *Bioinformatics*, 28(17), 2223–2230. doi:10.1093/bioinformatics/bts429
- Jones, K. R., Whitmire, J. M., & Merrell, D. S. (2010). A Tale of Two Toxins: *Helicobacter Pylori* CagA and VacA Modulate Host Pathways that Impact Disease. *Frontiers in microbiology*, 1, 115. doi:10.3389/fmicb.2010.00115
- Kelley, L. a, & Sternberg, M. J. E. (2009). Protein structure prediction on the Web: a case study using the Phyre server. *Nature protocols*, 4(3), 363–71. doi:10.1038/nprot.2009.2
- Kiran, M., Chauhan, A., Dziedzic, R., Maloney, E., Mukherji, S. K., Madiraju, M., & Rajagopalan, M. (2009). Mycobacterium tuberculosis ftsH expression in response to stress and viability. *Tuberculosis (Edinburgh, Scotland)*, 89 Suppl 1, S70–3. doi:10.1016/S1472-9792(09)70016-2
- Kuck, D., Kolmerer, B., Iking-Konert, C., Krammer, P. H., Stremmel, W., & Rudi, J. (2001). Vacuolating cytotoxin of *Helicobacter pylori* induces apoptosis in the human gastric epithelial cell line AGS. *Infection and immunity*, 69(8), 5080–7. doi:10.1128/IAI.69.8.5080-5087.2001
- Kutter, S., Buhrdorf, R., Haas, J., Schneider-Brachert, W., Haas, R., & Fischer, W. (2008). Protein subassemblies of the *Helicobacter pylori* Cag type IV secretion system revealed by localization and interaction studies. *Journal of bacteriology*, 190(6), 2161–71. doi:10.1128/JB.01341-07

- Lagesen, K., Hallin, P., Rødland, E. A., Staerfeldt, H.-H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic acids research*, 35(9), 3100–8. doi:10.1093/nar/gkm160
- Leekitcharoenphon, P., Lukjancenko, O., Friis, C., Aarestrup, F. M., & Ussery, D. W. (2012). Genomic variation in *Salmonella enterica* core genes for epidemiological typing. *BMC genomics*, 13, 88. doi:10.1186/1471-2164-13-88
- Lehours, P., Vale, F. F., Bjursell, M. K., Melefors, O., Advani, R., Glavas, S., & Guegueniat, J. (2011). Genome Sequencing Reveals a Phage in *Helicobacter pylori*, 2(6), 1–11. doi:10.1128/mBio.00239-11.Editor
- Lukjancenko, O., Ussery, D. W., & Wassenaar, T. M. (2012). Comparative genomics of *Bifidobacterium*, *Lactobacillus* and related probiotic genera. *Microbial ecology*, 63(3), 651–73. doi:10.1007/s00248-011-9948-y
- Mao, C., Shukla, M., Larrouy-Maumus, G., Dix, F. L., Kelley, L. a, Sternberg, M. J., Sobral, B. W., et al. (2013). Functional assignment of *Mycobacterium tuberculosis* proteome revealed by genome-scale fold-recognition. *Tuberculosis (Edinburgh, Scotland)*, 93(1), 40–6. doi:10.1016/j.tube.2012.11.008
- Mikkonen, T. P. (2004). Phylogenetic analysis of gastric and enterohepatic *Helicobacter* species based on partial HSP60 gene sequences. *International Journal of Systematic and Evolutionary Microbiology*, 54(3), 753–758. doi:10.1099/ijs.0.02839-0
- Moldavski, O., Levin-Kravets, O., Ziv, T., Adam, Z., & Prag, G. (2012). The hetero-hexameric nature of a chloroplast AAA+ FtsH protease contributes to its thermodynamic stability. *PloS one*, 7(4), e36008. doi:10.1371/journal.pone.0036008
- Morita, R., Nakane, S., Shimada, A., Inoue, M., Iino, H., Wakamatsu, T., Fukui, K., et al. (2010). Molecular mechanisms of the whole DNA repair system: a comparison of bacterial and eukaryotic systems. *Journal of nucleic acids*, 2010, 179594. doi:10.4061/2010/179594
- Mossey, P., Hudacek, A., & Das, A. (2010). *Agrobacterium tumefaciens* type IV secretion protein VirB3 is an inner membrane protein and requires VirB4, VirB7, and VirB8 for stabilization. *Journal of bacteriology*, 192(11), 2830–8. doi:10.1128/JB.01331-09
- Palframan, S. L., Kwok, T., & Gabriel, K. (2012). Vacuolating cytotoxin A (VacA), a key toxin for *Helicobacter pylori* pathogenesis. *Frontiers in cellular and infection microbiology*, 2, 92. doi:10.3389/fcimb.2012.00092
- Salama, N., Guillemin, K., McDaniel, T. K., Sherlock, G., Tompkins, L., & Falkow, S. (2000). A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. *Proceedings of the National Academy of Sciences of the United States of America*, 97(26), 14668–73. doi:10.1073/pnas.97.26.14668
- Salama, N. R., Gonzalez-Valencia, G., Deatherage, B., Aviles-Jimenez, F., Atherton, J. C., Graham, D. Y., & Torres, J. (2007). Genetic analysis of *Helicobacter pylori* strain populations colonizing the stomach at different times postinfection. *Journal of bacteriology*, 189(10), 3834–45. doi:10.1128/JB.01696-06

- Soares, S. C., Abreu, V. A. C., Ramos, R. T. J., Cerdeira, L., Silva, A., Baumbach, J., Trost, E., et al. (2012). PIPS: pathogenicity island prediction software. *PloS one*, 7(2), e30848. doi:10.1371/journal.pone.0030848
- Tatusov, R. L., Galperin, M. Y., Natale, D. A., & Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic acids research*, 28(1), 33–6. Retrieved from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102395&tool=pmcentrez&rendertype=abstract>
- Tomb, J. F., White, O., Kerlavage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D., Ketchum, K. A., et al. (1997). The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature*, 388(6642), 539–47. doi:10.1038/41483
- Walldén, K., Williams, R., Yan, J., Lian, P. W., Wang, L., Thalassinou, K., Orlova, E. V., et al. (2012). Structure of the VirB4 ATPase, alone and bound to the core complex of a type IV secretion system. *Proceedings of the National Academy of Sciences of the United States of America*, 109(28), 11348–53. doi:10.1073/pnas.1201428109
- Wu, S., Zhu, Z., Fu, L., Niu, B., & Li, W. (2011). WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC genomics*, 12(1), 444. doi:10.1186/1471-2164-12-444
- You, Y., He, L., Zhang, M., Fu, J., Gu, Y., Zhang, B., Tao, X., et al. (2012). Comparative Genomics of *Helicobacter pylori* Strains of China Associated with Different Clinical Outcome. *PloS one*, 7(6), e38528. doi:10.1371/journal.pone.0038528
- Zhang, R., & Lin, Y. (2009). DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic acids research*, 37(Database issue), D455–8. doi:10.1093/nar/gkn858
- Zhang, S., Moise, L., & Moss, S. F. (2011). *H. pylori* vaccines. (2011). *Human vaccines* 7:11, 1153–1157.
- Zheng, L.-L., Li, Y.-X., Ding, J., Guo, X.-K., Feng, K.-Y., Wang, Y.-J., Hu, L.-L., et al. (2012). A comparison of computational methods for identifying virulence factors. *PloS one*, 7(8), e42517. doi:10.1371/journal.pone.0042517

LEGENDS

Table 1. List of genomes used in this study. Genomic feature, statistics, accession numbers (chromosome + plasmid) and pangenomic data of genus *H.* (38 *H. pylori* and 8 non-*pylori* species).

Table 2. Biological categorization of candidate essential gene families

The *H. pylori* essential gene families analysed by **SLEP** pipeline for prediction of surface proteins. This is followed by **Vaxign** program for prediction of potential vaccine candidates. The 13 candidate GFs obtained, then analysed by **WebMGA** for functional annotation and prediction of COG IDs.

Figure 1. Phylogenomic tree and heatmap analyses of the genus *H.*

The complete genomes from the genus *H.* including 38 *H. pylori* and 8 non-*pylori* genomes were retrieved from the NCBI ftp site. Genomics comparisons between the variable content of all the genomes were plotted as percentages of similarity on the heatmap using program Gegenees (version 1.1.4). The percentage of similarities were used to generate a phylogenomic tree with program SplitsTree (version 4.12.6).

Figure 2. Whole genome/proteome pairwise alignment and comparative analysis.

The translated genomes of 38 *H. pylori* are analysed by BLASTp analysis. Pairwise comparisons across *H. pylori* proteomes are plotted in Blast Matrix. The shared proteome between any two *H. pylori* genomes and the percentage of similarities are calculated and shown in corresponding boxes, where the color intensity indicated the similarity. The diagonal row of rectangular boxes in the matrix illustrates the internal homology against its own proteome. The Scale given shows lowest and highest homology.

Figure 3. Core and Pan genome estimation for the genus, *pylori* and non-*pylori* species.

The figure demonstrates the core (1,185) and pan genome (2,825) of 38 *H. pylori* species (Table 1). On the left, the collective chart represents clusters (GFs) for each genome (green), pan genome (blue), and core genome (red). On the right, individual charts representing pattern of: A) *H. pylori* Core genome B) Pan genome C) Clusters (GFs) D) New GFs.

Figure 4. Core Genome of the *H. pylori* strains classification based on COG functional category.

Three set of core genomes: the genus composing the core genome of all the strains, core genome of non-*pylori* and core genome of the *Pylori* strains. Based on COG functional categories all core genomes are classified and assigned to “Metabolism”,

“Cellular Processes and Signaling”, Information Storage and Processing”, and some to category of “Poorly Characterized”.

Figure 5. Pathogenicity islands in *H. pylori* genomes (Pan-Heatmap).

Heatmap analysis demonstrate high degree of variability on most of the PAIs across all 38 genomes. Among the 22 predicted PAIs, only PiHps 2, 4, 14 and 15 are present in at least 50% of the strains.

Figure 6. Conserved and variable pathogenicity islands in *H. pylori*.

Putative pathogenicity islands predicted in *H. pylori*. The reference genome *H. pylori* 26695 is selected for analysis, all the genomes are aligned and phylogenetically related non-pathogenic organism *Wolinella succinogenes* DSM 1740 is also included. PiHps 2, 4, 14 and 15 were found conserved (A,B,C and D). PiHps 8, 13 and 9 (E, F and G) are variable PAIs among *H. pylori* genome been analysed.

Genome Analysed		Genome Statistics			Pangenome Analysis				
Genomes	Size bp	Proteins	Accession	Plasmid	% AT	NO. CL	New CL	PGFs	CGFs
H. pylori 2017	1548238	1503	NC_017374.1		60.6989	1493	1493	1493	1493
H. pylori 2018	1562832	1508	NC_017381.1		60.7065	1494	29	1522	1465
H. pylori 26695	1667867	1579	NC_000915.1		61.1243	1537	206	1725	1314
H. pylori 35A	1566655	1500	NC_017360.1		61.1329	1478	116	1836	1283
H. pylori 51	1589954	1515	NC_017382.1		61.2306	1492	90	1917	1275
H. pylori 52	1568826	1509	NC_017354.1		61.0591	1484	68	1978	1269
H. pylori 83	1617426	1559	NC_017375.1		61.2774	1522	94	2064	1263
H. pylori 908	1549666	1511	NC_017357.1		60.7012	1493	27	2086	1257
H. pylori B38	1576758	1502	NC_012973.1		60.8408	1457	62	2137	1231
H. pylori B8 6296	1680029	1579	NC_014256.1	NC_014257.1	61.2156	1538	62	2182	1230
H. pylori Cuz20	1635449	1536	NC_017358.1		61.1362	1506	78	2248	1222
H. pylori ELS37	1669876	1564	NC_017063.1	NC_017064.1	61.1235	1503	65	2295	1222
H. pylori F16	1575399	1508	NC_017368.1		61.1155	1460	31	2303	1222
H. pylori F30	1579693	1498	NC_017365.1	NC_017369.1	61.1975	1460	28	2315	1222
H. pylori F32	1581461	1506	NC_017366.1	NC_017370.1	61.1431	1467	26	2327	1220
H. pylori F57	1609006	1530	NC_017367.1		61.2725	1485	19	2337	1219
H. pylori G27	1663013	1577	NC_011333.1	NC_011334.1	61.1301	1524	59	2385	1216
H. pylori Gambia 9424	1712468	1589	NC_017371.1	NC_017364.1	60.8762	1530	33	2411	1215
H. pylori HPAG1	1605736	1508	NC_008086.1	NC_008087.1	60.9332	1450	33	2436	1216
H. pylori HUP-B14	1607584	1507	NC_017733.1	NC_017734.1	60.9579	1449	38	2464	1216
H. pylori India7	1675918	1572	NC_017372.1		61.1026	1513	40	2494	1213
H. pylori J99	1643831	1505	NC_000921.1		60.8113	1455	23	2506	1212
H. pylori Lithuania75	1640673	1555	NC_017362.1	NC_017363.1	61.1347	1514	58	2553	1207
H. pylori P12	1684038	1587	NC_011498.1	NC_011499.1	61.2138	1523	30	2570	1200
H. pylori PeCan18	1660685	1543	NC_017742.1		60.9816	1481	20	2577	1200
H. pylori PeCan4	1638269	1532	NC_014555.1	NC_014556.1	61.0894	1488	34	2604	1198
H. pylori Puno120	1637762	1532	NC_017378.1	NC_017377.1	61.0953	1480	34	2627	1199
H. pylori Puno135	1646139	1539	NC_017379.1		61.1769	1492	17	2640	1199

H. pylori SAfrica7	1679829	1573	NC_017361.1	NC_017373.1	61.5752	1499	56	2688	1200
H. pylori Sat464	1567570	1464	NC_017359.1	NC_017356.1	60.9083	1422	17	2694	1200
H. pylori Shi112	1663456	1561	NC_017741.1		61.2273	1490	17	2699	1200
H. pylori Shi169	1616909	1516	NC_017740.1		61.1362	1458	9	2699	1199
H. pylori Shi417	1665719	1545	NC_017739.1		61.2294	1490	16	2711	1198
H. pylori Shi470	1608548	1518	NC_010698.2		61.2294	1465	12	2715	1197
H. pylori SJM180	1658051	1524	NC_014560.1		61.1009	1461	18	2726	1197
H. pylori SNT49	1610830	1514	NC_017376.1	NC_017380.1	61.0046	1467	23	2741	1196
H. pylori v225d	1595604	1510	NC_017355.1	NC_017383.1	61.0556	1447	18	2750	1193
H. pylori XZ274	1656544	1701	NC_017926.1	NC_017919.1	61.4275	1598	100	2825	1185
H. acinonychisstr Sheeba	1557588	1547	NC_008229.1	NC_008230.1	61.8311	1477	125	2938	1161
H. bizzozeronii CIII-1	1807534	1877	NC_015674.1	NC_015670.1	53.8264	1807	1084	4022	661
H. cetorum MIT00-7128	1960111	1779	NC_017737.1	NC_017738.1	65.4654	1722	485	4476	665
H. cetorum MIT99-5656	1847790	1728	NC_017735.1	NC_017736.1	64.4632	1599	218	4665	672
H. cinaedi PAGU611	2101402	2143	NC_017761.1	NC_017762.1	61.4452	2104	1533	6188	450
H. felis ATCC_49179	1672681	1674	NC_014810.2		55.4913	1622	491	6616	465
H. hepaticus ATCC 51449	1799146	1802	NC_004917.1		64.0687	1784	514	7084	493
H. mustelae 12198	1578097	1426	NC_013949.1		57.5347	1404	690	7749	463

TABLE 1. *H. pylori* genomes analysed in this study.

Per. AT = Percentage AT; No. CL = Number of Clusters/GF; New CL = New cluster/GFs; PGFs = Pan gene families; CGFs = Core gene families

#COG	#	Annotation	Class	Class description
COG0671	1	Membrane-associated phospholipid phosphatase	I	Lipid transport and metabolism
COG0711	1	F0F1-type ATP synthase, subunit b	C	Energy production and conversion
COG0744	1	Penicillin-binding protein	M	Cell wall/membrane/envelope biogenesis
COG0757	1	3-dehydroquinate dehydratase II	E	Amino acid transport and metabolism
COG0848	3	Biopolymer transport protein	U	Intracellular trafficking, secretion, and vesicular transport
COG1452	1	Organic solvent tolerance protein OstA	M	Cell wall/membrane/envelope biogenesis
COG1519	1	3-deoxy-D-manno-octulosonic-acid transferase	M	Cell wall/membrane/envelope biogenesis
COG1862	1	Preprotein translocase subunit YajC	U	Intracellular trafficking, secretion, and vesicular transport
COG3736	1	Type IV secretory pathway, component VirB8	U	Intracellular trafficking, secretion, and vesicular transport
COG4736	1	Cbb3-type cytochrome oxidase, subunit 3	O	Posttranslational modification, protein turnover, chaperones

Tabl

e 2. Functional annotation of predicted vaccine candidate sequences.

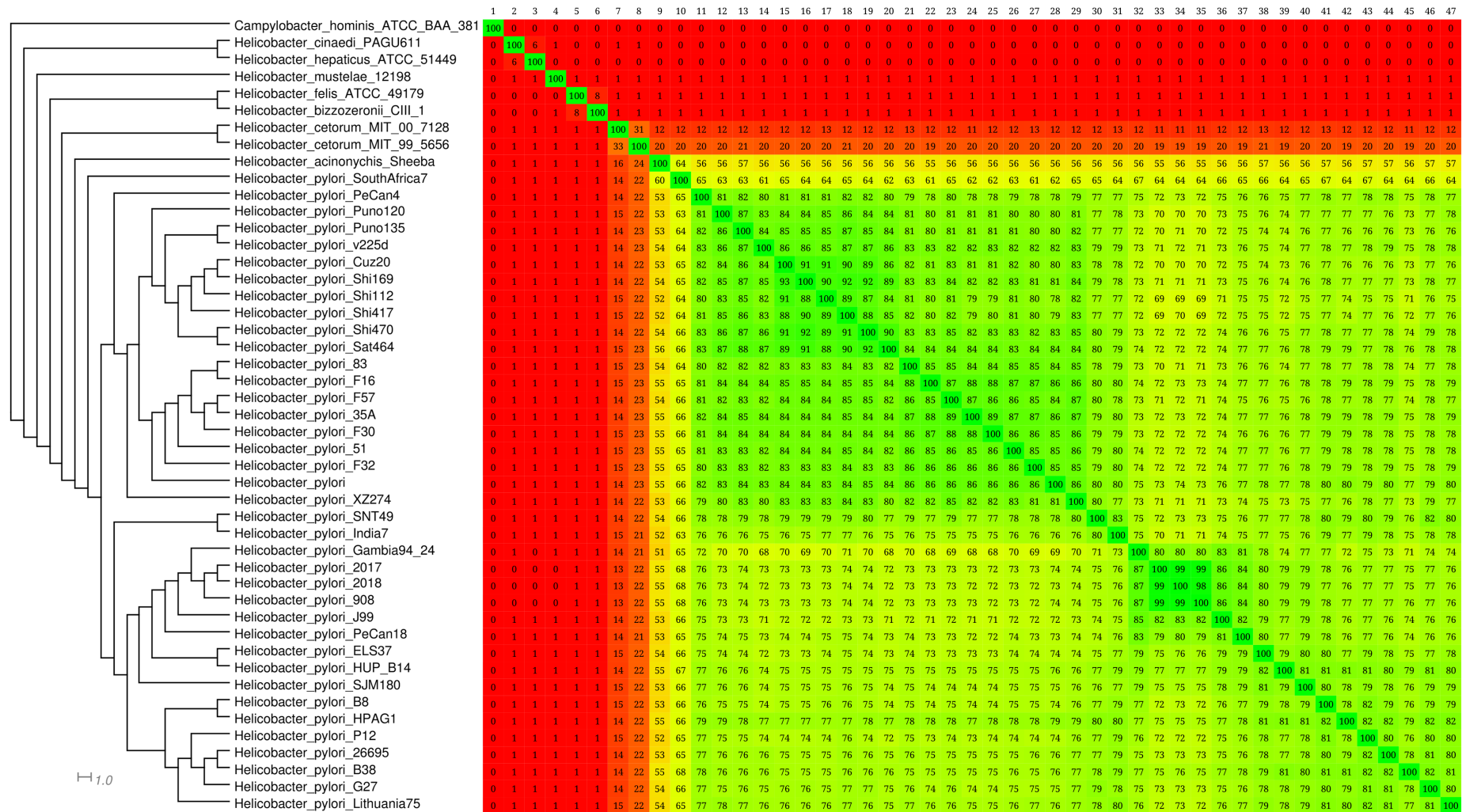


Figure 2.

Reliobacter_phot_2017										Reliobacter_phot_2018										Reliobacter_phot_20605										Reliobacter_phot_201										Reliobacter_phot_202										Reliobacter_phot_203										Reliobacter_phot_204										Reliobacter_phot_205										Reliobacter_phot_206										Reliobacter_phot_207										Reliobacter_phot_208										Reliobacter_phot_209										Reliobacter_phot_210										Reliobacter_phot_211										Reliobacter_phot_212										Reliobacter_phot_213										Reliobacter_phot_214										Reliobacter_phot_215										Reliobacter_phot_216										Reliobacter_phot_217										Reliobacter_phot_218										Reliobacter_phot_219										Reliobacter_phot_220										Reliobacter_phot_221										Reliobacter_phot_222										Reliobacter_phot_223										Reliobacter_phot_224										Reliobacter_phot_225										Reliobacter_phot_226										Reliobacter_phot_227										Reliobacter_phot_228										Reliobacter_phot_229										Reliobacter_phot_230										Reliobacter_phot_231										Reliobacter_phot_232										Reliobacter_phot_233										Reliobacter_phot_234										Reliobacter_phot_235										Reliobacter_phot_236										Reliobacter_phot_237										Reliobacter_phot_238										Reliobacter_phot_239										Reliobacter_phot_240										Reliobacter_phot_241										Reliobacter_phot_242										Reliobacter_phot_243										Reliobacter_phot_244										Reliobacter_phot_245										Reliobacter_phot_246										Reliobacter_phot_247										Reliobacter_phot_248										Reliobacter_phot_249										Reliobacter_phot_250										Reliobacter_phot_251										Reliobacter_phot_252										Reliobacter_phot_253										Reliobacter_phot_254										Reliobacter_phot_255										Reliobacter_phot_256										Reliobacter_phot_257										Reliobacter_phot_258										Reliobacter_phot_259										Reliobacter_phot_260										Reliobacter_phot_261										Reliobacter_phot_262										Reliobacter_phot_263										Reliobacter_phot_264										Reliobacter_phot_265										Reliobacter_phot_266										Reliobacter_phot_267										Reliobacter_phot_268										Reliobacter_phot_269										Reliobacter_phot_270										Reliobacter_phot_271										Reliobacter_phot_272										Reliobacter_phot_273										Reliobacter_phot_274										Reliobacter_phot_275										Reliobacter_phot_276										Reliobacter_phot_277										Reliobacter_phot_278										Reliobacter_phot_279										Reliobacter_phot_280										Reliobacter_phot_281										Reliobacter_phot_282										Reliobacter_phot_283										Reliobacter_phot_284										Reliobacter_phot_285										Reliobacter_phot_286										Reliobacter_phot_287										Reliobacter_phot_288										Reliobacter_phot_289										Reliobacter_phot_290										Reliobacter_phot_291										Reliobacter_phot_292										Reliobacter_phot_293										Reliobacter_phot_294										Reliobacter_phot_295										Reliobacter_phot_296										Reliobacter_phot_297										Reliobacter_phot_298										Reliobacter_phot_299										Reliobacter_phot_300										Reliobacter_phot_301										Reliobacter_phot_302										Reliobacter_phot_303										Reliobacter_phot_304										Reliobacter_phot_305										Reliobacter_phot_306										Reliobacter_phot_307									
89.9	90.0	89.9	89.9	91.0	91.0	91.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0	92.0	91.0																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																

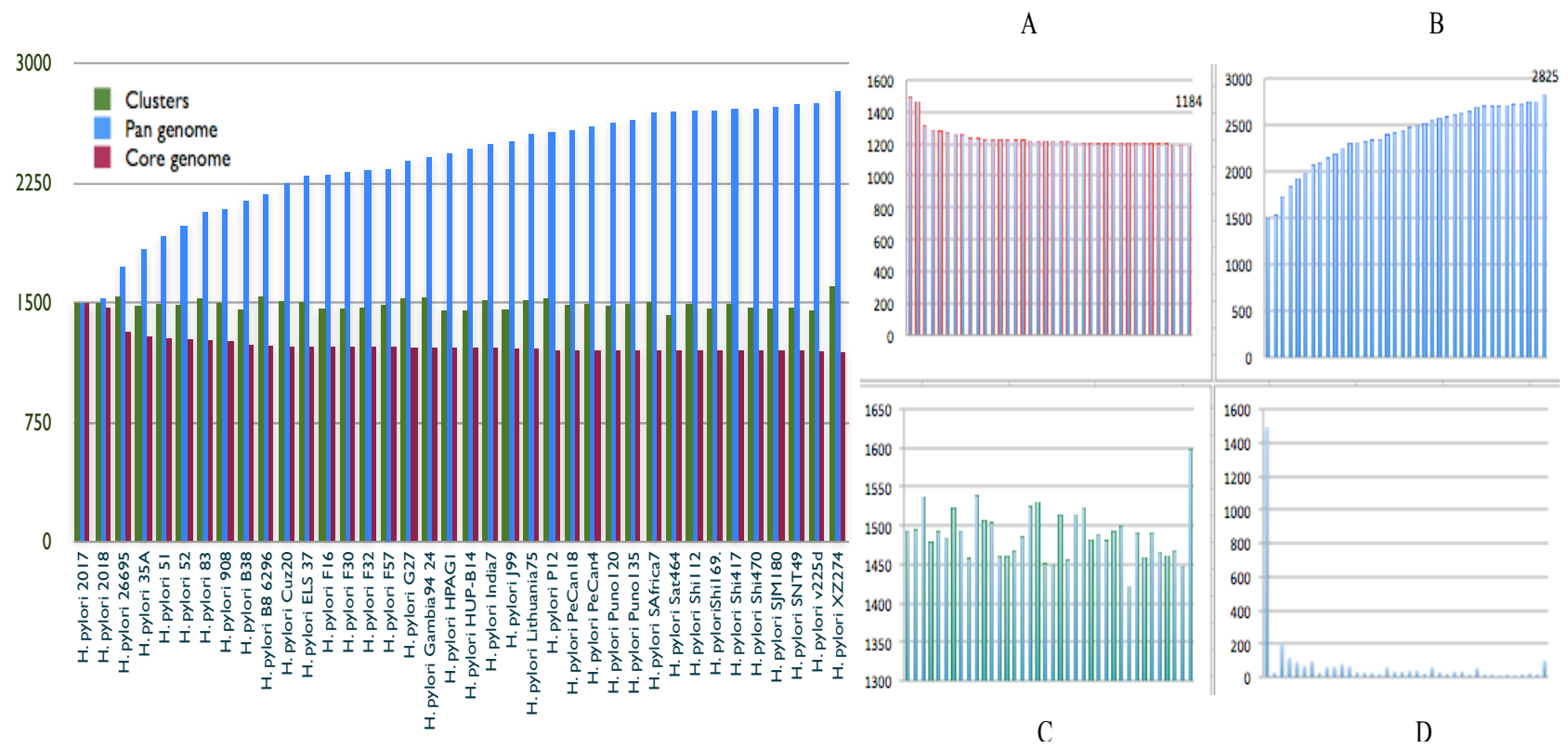


Figure 3.

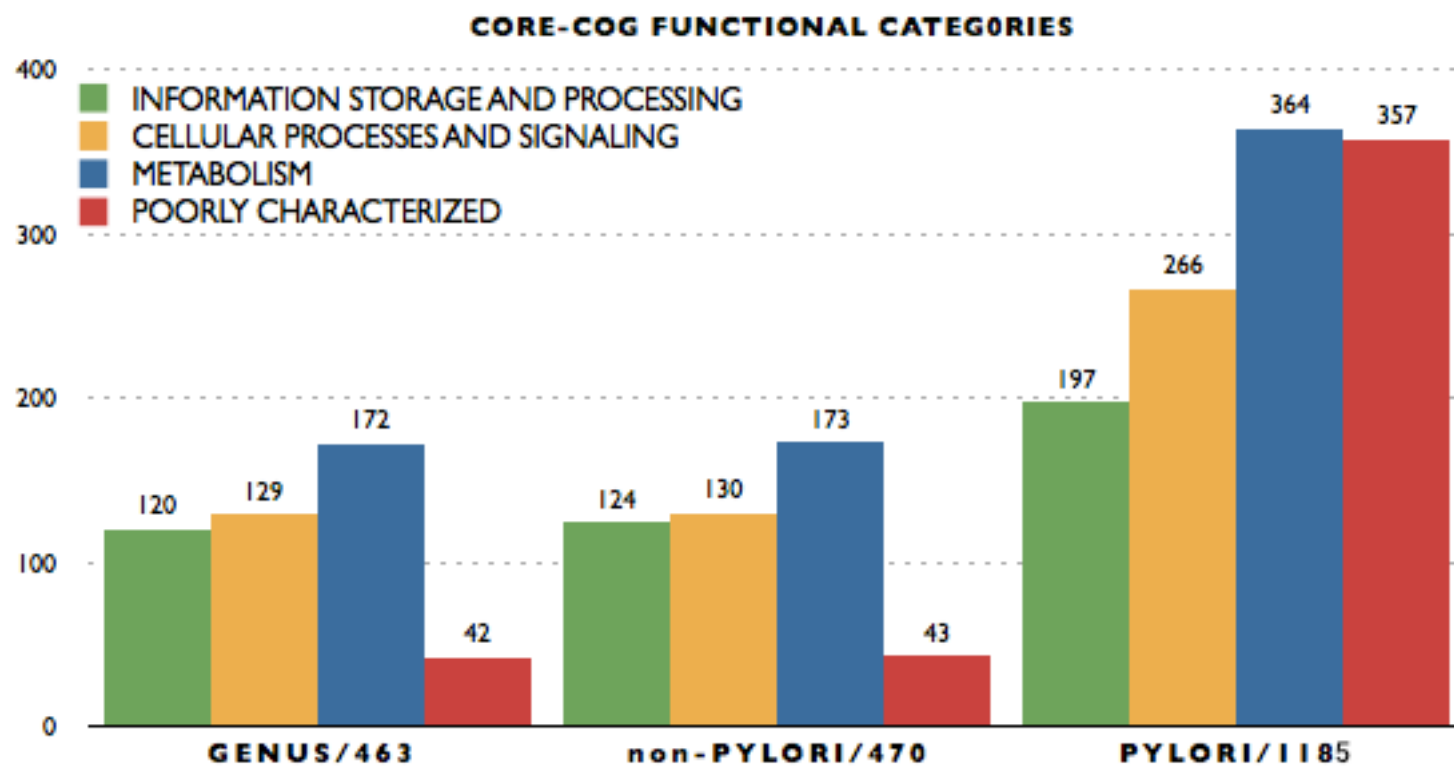


Figure 4.

Figure 5.

Strains	26695PIHp1	26695PIHp2	26695PIHp3	26695PIHp4	26695PIHp5	26695PIHp6	26695PIHp7	26695PIHp8	26695PIHp9	26695PIHp10	26695PIHp11	26695PIHp12	26695PIHp13	26695PIHp14	26695PIHp15	26695PIHp16	26695PIHp17	26695PIHp18	Pecan4PIHp22	Cuz20PIHp19	SouthAfrica7PIHp20	SouthAfrica7PIHp21
SouthAfrica7	92	100	95	88	100	68	94	29	24	82	94	77	43	100	100	89	79	87	81	60	100	100
Pecan4	85	100	95	100	94	77	88	29	97	82	81	77	43	100	100	78	86	87	100	51	45	83
Puno120	90	94	90	88	94	77	76	25	91	91	94	77	33	100	100	67	79	87	76	33	45	54
Puno135	95	100	90	100	94	77	82	29	97	91	94	77	48	100	100	89	86	80	81	51	45	85
v225d	92	94	86	100	88	77	82	32	97	91	88	85	14	91	100	89	79	87	81	31	55	54
Cuz20	85	94	90	88	94	82	88	7	94	82	94	92	43	91	100	67	86	87	81	100	45	98
Shi169	83	100	95	100	94	82	88	7	94	91	94	69	38	91	100	89	86	80	81	60	45	98
Shi112	90	100	95	100	94	64	82	11	91	91	81	92	48	100	100	67	79	87	81	72	36	96
Shi417	90	100	95	88	94	82	88	36	94	82	88	85	52	91	100	78	93	93	81	58	45	93
Shi470	87	100	95	100	88	82	82	7	97	73	88	85	43	100	100	89	79	87	81	63	45	93
Sat464	87	94	95	100	94	82	82	4	70	82	88	77	24	100	100	78	86	87	81	31	45	52
83	92	94	95	100	100	86	94	11	100	100	94	85	33	100	100	89	93	93	100	38	45	59
F16	92	100	95	88	94	77	88	0	97	91	94	85	14	100	100	89	93	93	81	41	55	54
F57	92	94	95	88	94	82	88	25	94	91	94	77	43	100	100	89	86	93	81	62	55	100
35A	85	94	81	88	94	77	76	14	100	100	88	62	10	100	100	67	79	93	81	31	55	54
F30	92	100	95	88	94	82	82	14	100	91	94	92	14	100	100	89	79	87	81	31	45	54
51	83	94	95	88	94	73	88	7	100	73	94	62	38	91	100	89	79	93	81	36	55	63
F32	90	100	95	88	94	77	88	11	97	82	94	85	38	100	100	89	93	93	81	33	55	52
52	92	94	95	100	94	73	88	0	100	82	88	85	10	100	100	89	79	93	76	28	45	54
XZ274	83	89	81	75	71	55	88	32	88	73	81	62	48	100	100	78	79	67	76	55	55	93
SNT49	85	94	95	100	100	64	76	7	97	73	94	85	52	100	100	56	86	73	81	59	55	96
India7	92	94	95	100	94	64	82	11	94	82	88	77	52	100	100	89	86	87	81	70	55	57
Gambia94/24	92	94	90	100	94	64	82	11	94	82	88	69	62	100	100	89	86	87	81	74	55	80
2017	90	100	90	100	94	64	100	0	100	82	81	62	19	82	100	89	93	53	81	43	55	74
2018	93	100	95	100	94	64	94	0	100	82	81	85	19	82	100	89	93	53	81	44	55	74
908	93	94	95	100	88	64	94	4	97	82	81	62	19	82	83	89	93	53	81	43	55	72
J99	92	100	95	100	88	59	94	0	100	82	94	85	33	100	100	100	93	93	81	38	55	61
Pecan18	90	100	95	75	94	64	82	11	97	82	88	85	43	100	100	78	71	73	81	34	55	52
ELS37	92	100	86	100	94	64	82	11	94	73	81	85	33	100	100	78	86	93	81	37	73	57
HUP-B14	92	89	95	100	94	64	76	0	91	82	88	77	29	100	100	78	79	87	81	30	73	50
SJM180	92	100	90	100	94	64	82	0	97	82	94	54	43	100	100	78	86	100	81	38	64	59
B8	95	94	95	100	94	59	88	32	100	91	94	92	62	100	100	100	93	87	81	47	91	72
HPAG1	93	100	95	100	94	64	82	0	100	91	94	85	10	100	100	89	93	93	81	31	55	54
P12	98	94	90	100	94	82	88	57	100	91	94	92	71	100	100	78	93	87	81	50	55	78
26695	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	81	40	55	61
B38	88	89	86	100	94	68	88	21	24	73	88	77	19	91	100	89	86	87	81	33	55	52
G27	90	94	95	100	94	77	88	25	94	82	94	85	57	91	100	89	86	80	81	57	55	91
Lithuania75	82	100	90	88	94	77	88	18	97	73	94	69	33	100	100	67	93	100	81	35	45	52



3.2 Conclusion

Considering *H. pylori* as a potential human gastric pathogen and its worldwide distribution, limited attention has been given to explore the genomic information of this pathogen. We therefore, for the first time attempted to exploit the larger set of *H. pylori* genomic data to estimate the intra-species relationships, phylogenomics and common and unique feature associated with species. We were able to compare each proteome (genome) with the others in the data set and calculate the number of conserved proteome between and among species. It was found that the shared proteome was considerably high and also higher number of genes were observed in the core genome for 38 *H. pylori* species. The species core essential genes were predicted consisting of 493 gene families which were then analysed for therapeutic candidates identifications. The use of various immunoinformatic strategies resulted in filtering target sequences for broad range vaccine designed against *H. pylori*. We hope the data generated during this study will assist researcher to understand the behavior of the organism and mechanisms of pathogenicity. Moreover the essential genes sequences which are vital to the pathogen and required for survival and predicted in this study will provide assistance in development of antibacterial therapeutics (drugs).

V. DISCUSSION

In this project, all the analysed complete genome sequences (species) were downloaded from NCBI genbank along with their published annotations. Almost all of the genomes had annotation except *Campylobacter fetus* subspecies *venerealis*. As different groups had been involved in annotation of these genomes with their own standards, therefore, no uniform standard operation procedures had been followed. Moreover, different programs have been used for ORFs (genes) prediction. Consequently, some irregularities can be observed in the genomes deposited to genbank (number of ORFs, tRNA, rRNA and annotation). To avoid confusions and inconsistencies in the data, all the genbank (.gbk) files were downloaded and DNA sequences were extracted in fasta formats. Then, they were subjected to a single program “Prodigal”, a gene finding algorithm to predict ORFs (genes). The program has already been tested on various bacteria (including, *Corynebacterium*, *Campylobacter*, and *Helicobacter* species) and some archaea genomes for prediction of genes with optimum specificity and sensitivity (<http://prodigal.ornl.gov/results.php>).

This strategy helped in uniformity of resultant data. When compared to the data of public submitted statistics (ORFs), our predicted genes (in each genome) were observed with slight variation but the number of genes were sufficiently comparable. This analysis also highlighted the issue that there is an urgent need of standardization and uniformity in gene finding and annotation strategies (Ali et al., 2013; Azevedo et al., 2011; Hyatt, LoCascio, Hauser, & Uberbacher, 2012).

Before we discuss our results, it is important to define the differences between the terms homology and similarity. Homology implies common ancestry of two genes or gene products. Similarity is what we can measure from alignment of sequences or structures. Similarity may be used as evidence for homology, but does not necessarily imply homology. Most of the studies described in this manuscript are based on analysis of similarity, which is a means towards studying homology.

In comparative proteome (proteins) analysis, results shown in the blast matrices are obtained by bi-directional (reciprocal: A points to B and vice versa) whole genome (proteome) comparisons and should not be confused with one way alignments. In this sense, if a proteome A is aligned with proteome B (of different size), the percentage of proteins shared by proteome A with proteome B is not necessarily the same as that of B sharing with A. As an example, the number of shared proteome of A with B is converted into percentage derived from the shared proteins between the two genomes (A shares X proteins with B) divided by total number of second genome (X/B). A detailed look into the matrix can explain the differences/similarities between and among species and genomes.

In Chapter 1, Our intention is to provide an overview of the genomic tools and developments in genomic science in the recent past. As genomics is a rapid evolving area of science, especially in the last few years great advancement has been made in NGS technology and associated bioinformatics tools. Majority of tools mentioned are quite updated, if not all, and are frequently used by the genomic community for similar studies. However, as sequencing technology is evolving rapidly it is suggested to the readers to consult the latest literature every time, especially for development and advances in sequencing technologies, assembly and alignment tools (Forde & O'Toole, 2013; Hughes, Gang, Murphy, Higgins, & Teeling, 2013).

It is important to note that previously taxonomic analyses were performed using a diverse and often arbitrary selection of morphological and phenotypic characteristics. Nowadays, these characteristics are generally considered unsuitable for generating reliable and consistent taxonomies for prokaryotes, because there is no rational basis for selecting which morphological or phenotypic properties should be tested. Additionally, it is doubtful that individual phenotypes or small collections of phenotypes can always correctly represent evolutionary relationships. Recently, the advent of DNA sequencing facilitated sequencing of genomic sequences and 16S rRNA gene sequence comparisons are becoming the standard

technique for taxonomic analyses (Lagesen et al., 2007; B. Trost et al., 2010).

Based on the notion that most of the evolutionary relationships between microorganisms are inferred by comparison of single genes, usually 16S rRNA genes, the same approach has been considered as the preliminary analysis in this project too. Although extremely effective, single gene phylogenetic trees only provide limited information which can make determining broad relationships between major groups difficult. Therefore, phylogenetic relationships have also been determined by whole genome comparisons using the observed absence or presence of protein coding gene families in group of organisms (refer to chapter 1, figure 5A and B; chapter 2, Figure 1 and 3B). In fact this is similar to using the distribution of morphological characteristics to determine phylogenetics without the problem of convergent evolution. Trees produced using pangenome (pangenome) methods are similar to 16S rRNA trees. As more genome sequences become available more detailed conclusions can be drawn using the mentioned method (Agren, Sundström, Håfström, & Segerman, 2012; Jolley et al., 2012; B. Trost et al., 2010).

One step ahead, test results for both phylogenetic tree and genomic variation analysis based on whole genomes sequences have been discussed in chapter 3 (Figure 1). An alternative strategy, phylogenomic (heatmap) analysis was examined in which the whole genomes are first broken into pieces and then individual pieces are aligned to estimate the similarities in genomic contents corresponding to phylogenetic tree derived from the data. The results obtained in this analysis were found interesting when compared to proteomic comparisons of the same organisms genomes, which are given in Figure 2 of the same chapter. The distribution of the genomes In both analysis (genomic variation and proteome comparison) resembles to each other and to that of the phylogenetic tree (Appendix, additional figure 1 of chapter 3). This suggests the applicability of the strategies to understand the genomic relationships and their position in the group (species or genus).

Our analysis and observations also indicated that both 16S rRNA (single gene) phylogenetic analysis and the pangenomic tree (presence or absence of specific genes in species) are equally interesting and helpful in classification and characterization of species or genus (Snipen & Ussery, 2010). This suggests that both strategies could be applied to better understand the position of specific genome in the phylogenetic tree.

By parallel pangenome analysis of different species (all the three genera), it should be mentioned with confidence that high pangenome variation at genus (*Corynebacterium*, *Campylobacter* and *Helicobacter*) level reflects the considerable diversity in species and adaptation to diverse environments. For example, the diversity in *Corynebacterium* species can be seen as they are taxonomically classified into diverse nature and environments including industrially important bacteria, commensals of animals as well as humans pathogens and, hence, reflected by the analyses (Ott *et al.*, 2012).

In general, the species from the genus *Corynebacterium* have high G+C content in their genomes. However, in one member of the genus *C. pseudotuberculosis* species it was observed that their genomes have lower GC content (52%) and reduced gene repertoires (Av. genes 2,145) indicating that the species have lost numerous genes. It has also been demonstrated previously that a high degree of conservation of gene order exists in four *Corynebacterium* genomes, *C. diphtheriae*, *C. glutamicum*, *C. efficiens* and *C. jeikeium*, showing only 10 gene-order breakpoints. Also, rearrangement events during evolution in these species appear to be rare (Ruiz *et al.*, 2011). This suggests that pathogenic species are more genetically conserved in the genus. Similar conservation has also been observed in translated genomes (proteome) comparisons.

From pathogenomic point of view, the three *Corynebacterium* species: *C. diphtheriae*, *C. ulcerans* and *C. pseudotuberculosis* shared some common pathogenic life style and on the gene levels analysis it has been explained that the main virulence factor common to them is the diphtheria toxin. The reason of diphtheria-like illness caused by *C. ulcerans* may be due

to harbouring lysogenic β -corynephages, which codes the diphtheria toxin and is responsible for the systemic symptoms caused by *C. ulcerans* while producing diphtheria toxin, similar to that encoded by toxigenic strains of *C. diphtheriae*. On the other side, *C. pseudotuberculosis* can also harbour diphtheria “tox genes” and express diphtheria toxin. Infections by these species most often occur in animals but can also cause diphtheria-like disease in humans, with symptoms similar to those caused by *C. diphtheriae* (Buck, Cross, Wong, Loera, & Groman, 1985; Groman et al., 1984; Wagner et al., 2010).

Beside the individual genomes (proteome) comparison for similar proteins, pangenome is a better strategy for generalizing the overall picture of conservation in related species (strains). Based on our observation, an intra-species (strains from the same species) pangenome analysis results in higher conserved genomic sequences, while inter-species may show significant diversity. For example thirteen *C. diphtheriae* strains found to contain 1,632 conserved genes in them (Trost *et al.*, 2012) and similarly, fifteen *C. pseudotuberculosis* strains analysis showed core genome consisting of 1,504 genes (Soares et al., 2013; E. Trost et al., 2012), about 68% of the average genome (2195 genes). In contrast to intra-species comparisons, we noticed a huge diversity in inter-species comparison, for example, the 11 species of *Corynebacterium* were found to share 741 genes families, which is about 31%, of the average genome size (2,391) in the selected species. On the other side, even intra-species differences show that *C. pseudotuberculosis* biovar ovis is more clonal than biovar equi, and its pangenome grows at slow rate. It presents an even bigger core genome due to higher clonality, while inter-species comparison indicated that *C. pseudotuberculosis* is more clonal than *C. diphtheriae* and the pangenome grows at slow rates, possibly due to facultative intracellular lifestyle (Soares *et al.*, 2013).

In contrast, lower genomic conservation in *Corynebacterium* inter-species (genus) supports the theory that even species from the same genus have greater variations resulting in considerable diversity (Ali *et al.*, 2013). It is also proposed by this study that with the increase

of other sequenced species from the genus, the pangenome of the genus will increase with almost similar rate and the core genome might decrease but at relative low rate, this is an indication of the inter-species open pangenome (Lapierre & Gogarten, 2009).

Our analyses of the genomes plasticity in *Corynebacterium* species, *C. kroppenstedtii*, a pathogenic species is selected as a reference organism, the analyses highlighted the genomic variations and deletions pattern in all other *Corynebacterium* species. The specific regions i.e. predicted pathogenicity islands (PAIs) in *C. kroppenstedtii* present high concentration of hypothetical genes. However, they will account for the singletons of this species and can be related to new functions and adaptability to new environments/hosts. Furthermore, if these genes are characterized they may contribute to the pathogenomics of the species. The significance of this model analysis could be more observable from our previous results. We previously analysed another species *C. pseudotuberculosis* from the same genus and the number of the PAIs observed in two biovars of *C. pseudotuberculosis* strains (15) were 16. Among the interesting results with respect to the gene content of the PAIs, we observed high similarity of the pilus genes in the biovar ovis strains compared with the great variability of these genes in the biovar equi strains (Ruiz *et al.*, 2011; Soares *et al.*, 2013). Moreover, the polymerization of complete pilus structures in biovar ovis is suspected to be responsible for a significant ability of these strains to spread through host tissues and for penetrating cells to live intracellularly. With these findings we propose that, it may be interesting if other pathogenic species from the genus are analysed for their genomic regions associated with mechanism of pathogenicity. The next target in the genus could be *C. ulcerans*, detected as a commensal in domestic and wild animals and probably serving as reservoirs for zoonotic infections. However, recently, the rate and severity of human infections (diphtheria-like illnesses) with *C. ulcerans* seem to be increasing in some countries. Hence, *C. ulcerans* could be a suitable target for the same analysis (Wagner *et al.*, 2010).

In chapter 2, we presented comparative genomic and proteomic analysis of different species from the genus *Campylobacter*. In general, *Campylobacter* species demonstrate significant diversity both on taxonomic and pathologic levels. They are highly adapted to mucosal surfaces and associated with human and/or animal infections ((Moolhuijzen *et al.*, 2009; Ali *et al.*, 2012). Our interests in the genus is because of the diverse pathogenic nature of the species and their wide host range. Secondly, our group recently published a complete genome sequence of *C. fetus* subsp. *venerealis*, a veterinary pathogen (Stynen *et al.*, 2011). Among species in genus, *C. fetus* subspecies: *C. fetus* subsp. *venerealis* and *C. fetus* subsp. *fetus* show clonal population structure and are targeted for more detailed analysis.

The phylogenetic analysis based on ribosomal RNA genes sequences divergence resulted in *C. fetus* subspecies central position in the genus, almost equally close to *C. jejuni* species and subspecies as well as and non-*jejuni* species (*C. lari*, *C. curvus* etc.). However, it should not be related to their genomic similarity because an overall significant diversity has been observed in inter-species (Ali *et al.*, 2012; Moolhuijzen *et al.*, 2009). An interesting finding in the tree is that *C. fetus* subspecies are positioned close to *C. hominis* species, the only known non-pathogenic species in the genus *Campylobacter*. This was also the reason to select this genome (reference non-pathogenic) for both comparative pathogenomic analysis and genome plasticity analysis in both the *C. fetus* subspecies, which are also the part of this project.

Despite the high level of observed genetic relatedness (similarities) in their proteomic content (76%), the *C. fetus* subspecies (*Cff* and *Cfv*) exhibit distinct host tissue preferences (Gorkiewicz *et al.*, 2010). The subspecies *venerealis* is largely restricted to the bovine genital tract, causing epidemic abortion in these animals and substantial economic losses to the cattle industry. *C. fetus* subsp. *fetus* is an important agent in ovine abortion worldwide and can also induce severe disease in humans. On the other side, 18 species comprising the

genus *Campylobacter* show similar pattern of host tissue specificity, which makes them excellent models to study their unique features and host-pathogen relationships (Gorkiewicz et al., 2010; Tu, Gaudreau, & Blaser, 2005).

Although the major focus was on *C. fetus subspecies* but *C. jejuni* species were also included in analysis to understand the genomic conservation and relationships between species in the genus. As it can be seen from figure 2, chapter 2, a dense green pyramid indicates the similarities in the *C. jejuni* and subsp. *jejuni* proteomes. The highest number of proteins shared are ~93% by *C. jejuni* subsp. *jejuni* 81116 with *C. jejuni* sb. *jejuni* M1 (1546/1653). However, in *C. jejuni* subspecies, as a whole, no significant genomic conservation was observed (except for few of them), in contrast to our expectation to have more conserved genomes. This may be due to the reason that they colonize the gastrointestinal tract of many (different) warm-blooded animals such as chickens, cattle, cats and dogs etc. similarly, multilocus sequence typing (MLST) data analysis has shown that the majority of human infections can be attributed to poultry (57–80%) and livestock (18–39%), while environmental exposure has only been recognized to play a minor role (1–4% of the human infections).

We were able to estimate the genomic diversity in genus which is clearly indicated by the pangenomic analysis. It was observed that the core genome (15 genomes) contains 552 gene families which is relatively low and the pan-genome comprises 7059 gene families, representative of the genomic variability in inter-species. As *Campylobacter* species have small genomes compared to other species, it was expected to contain a larger fraction of genes in the core genome and relatively fewer in the accessory genome (Snipen & Ussery, 2010). However, contradictory to our expectations, the core genome calculated is about 31% of an average genome (1,794 genes).

If we compare these results with that of genus *Corynebacterium* in chapter 1, the percentage is almost the same (~31%). However, the number of genome analysed were more (19) in

Corynebacterium and therefore, the high diversity in *Campylobacter* species became evident (15 genomes).

Despite higher genomic and proteomic (76%) similarity in *C. fetus* subspecies, it was surprising to observe that higher number of strain specific genes, for example, 428 unique genes were found associated with *C. fetus* subsp. *venerealis*. This is a relatively higher number than expected and 88 gene families were found specific to *C. fetus* subsp. *fetus*. We proposed two possible explanations to this higher number of unique genes in *C. fetus* subsp. *venerealis*. First, the organism has the biggest genome size (relatively bigger 1.8 Mb) in the genus and therefore more ORFs in genome. Secondly, the organism is confined to specific host tissue and lifestyle, therefore most probably has acquired DNA from neighbor organisms. In contrast to *C. fetus* subsp. *venerealis*, other comparative genomic analyses of *Campylobacter* species have revealed a process of genome decay supported by a small genome size (about 1.5 Mb) and the loss of metabolic genes consistent with successful adaptation to a specific niche. It has also been explained that *Campylobacter* genomes are among the most dense bacterial genomes known, with about 95% coding sequence. Nevertheless, evidence of reduction and plasticity in genetic composition remain apparent, as strain-specific genes comprise a substantial proportion of the entire repertoire of 1,500 to 1,800 genes (Gorkiewicz *et al.*, 2010). As expected, in functionally characterized core genome of the genus, majority of the genes/proteins are observed to have some role in basic cellular processes e.g. metabolic and biosynthesis. Beside the core genome, strain-specific genes predicted in this study require special attention. If functionally characterized, they might give some clues about the subspecies host-specificity, clonal behavior and mechanism of disease in particular host (Kienesberger *et al.*, 2011).

Proteins subcellular localization in *C. fetus* subspecies was one of the important parts of this project, with the aim to discover potential core immunogenic and vaccine candidate proteins. As noticed in the post genomic era, the prediction of protein localization in and out of the

bacterial cell is of much importance for exploitation of the overwhelming amounts of DNA sequences data for identification of potential immunogens. The growing interest in certain bacterial proteins is getting higher. For example, cell surface proteins and exported (secreted) proteins are important for their immunogenic properties, which may serve as targets for vaccine development. This is due to the fact that they mediate physical interactions between pathogen and host and consequently play a role in activation of host responses (Barinov et al., 2009; Giombini, Orsini, Carrabino, & Tramontano, 2010). There are different programs and pipelines available for this propose including, SurfG+ (Barinov et al., 2009), Vaxign (He et al., 2010), PSORTb (Gardy et al., 2005), SLEP (surface proteins predictor) (Giombini et al., 2010) etc. The pipeline SurfG+ is used for proteins subcellular localization in *Campylobacter* species because of the permission to alter setting in the program. Also, unlike other programs to predict surface (membrane) and secreted (SEC) proteins, potential surface exposed (PSE) can also be localized. The core-exoproteome predicted by SurfG+ contains 285 proteins sequences (127 SEC and 158 PSE), which were further divided into groups and the best top group was selected. It consisted of 77 potential immunogenic candidate sequences. These sequence were crosschecked for validation by Vaxign tool. Among the 77 sequences, 4 sequences were predicted as cytoplasmic were therefore disregarded. However, it must be notified that they could be used as drug targets. The final set of 73 potential immunogenic sequences were subjected to blast2go program for functional annotation. As it was expected, most of the candidates have functions in metabolic and biosynthesis processes, in addition to these few sequences were found to be involved in pathogen growth, survival and pathogenesis. These sequences are predicted to be potential immunogenic and can contribute in the development of vaccines against *C. fetus* subspecies and are available for public use and can be download by visiting the following link <http://dx.doi.org/10.1016/j.gene.2012.07.070> (Ali et al., 2012).

In figure 5 of the same chapter, a technical error took place where *Cj* 1221 was shown to

have only 12 secreted proteins, whereas, the corrected number is 121 (crosschecked). For convenience in interpretation of the chart data, an additional table is provided in the appendix IX.II. It can also be noticed that the size of the *C. fetus subsp. venerealis* in base pairs in Table 1 (1,874,244 bp) is different from that of figure 4A (1,858,789 bp). This should not be consider incorrect. As there are two possible reasons for this, for all the genomes in this study the genome statistics (ORFs, GC, AT, RNA counts etc.) including size in base pairs are calculated (single strategy), while the complete genome sequence for pathogenicity analysis is directly obtained from genbank (as a requirement by the program PIPS). Additionally, the difference might also be because of the genome sequence availability in more than one contigs (Stynen *et al.*, 2011).

Reverse vaccinology approach applied in this study resulted in interesting findings. During the characterization of the core exo-proteome of *C. fetus* subspecies, a number of virulence factors and vaccine candidates genes/proteins were identified. Potential immunogenic sequences contain flagella associated proteins (FlgK, FlgE, FlgL). These were found conserved in *C. fetus* subspecies and are well documented virulence factors. They facilitate the colonization of pathogen and are associated with the processes of chemotaxis and motility (Terashima *et al.*, 2008; Moolhuijzen *et al.*, 2009). Comparative to other species, among all the predicted virulence factors, these factors and candidate proteins are found to be highly conserved across the genus (Moolhuijzen *et al.*, 2009).

Our analysis also resulted in the discovery of candidate nucleoside diphosphate kinase (Ndk), which is involved in essential biological processes and has significant role in bacterial growth, signal transduction and pathogenicity. Hence, it is presented as a good target for vaccine development against *C. fetus* subspecies. Ndk is also being reported in other pathogens such as *Pseudomonas aeruginosa*, *Trichenella spiralis*, *Vibrio cholera* and *Mycobacterium bovis* (BCG) (Chakrabarty, 1998; Narayanan & Ramaswami, 2003). The target Ndk is well studied in mycobacteria and it demonstrated the relevance of Ndk-

mediated deactivation of Rab5 and Rab7 to the virulence of mycobacteria by knocking-down Ndk gene expression in BCG, using antisense strategy. Increased fusion of phagosomes containing BCG AS-Ndk with lysosomes was observed along with a significant decrease in bacterial survival within the macrophage. This evidence supported the hypothesis that mycobacterial Ndk is a putative virulence factor that inhibits phagosome maturation and promotes survival of mycobacteria within the macrophage (Sun *et al.*, 2010).

Other vaccine target sequences identified in this project include surface layer proteins (S-layer) and surface array proteins (SAP) of *C. fetus* subspecies (Fogg, Yang, Wang, & Blaser, 1990; Umelo-Njaka *et al.*, 2001). From structural point of view, these proteins have unique features on surface i.e. the outermost crystalline layer is made of monomolecular proteins called S-layer proteins (SLP). These SLPs are important factors which mediate physical interactions between pathogen and host and are the key factors providing resistance to host immune defenses. They also facilitate virulence and therefore are considered as significant immunogenic targets and vaccine candidate (Blaser *et al.*, 1994; Dworkin, Shedd, & Blaser, 1997; Tu *et al.*, 2005). In the development of vaccine, S-layer has significant importance as it is the surface display proteins and it has epitope capability. These properties make these proteins excellent candidate for carrier antigens. In *Caulobacter crescent*, the S-layer is composed of a single protein, RsaA, and is among the most abundant encoded proteins. *Caulobacter's* RsaA protein was adapted to make 2 recombinant proteins of the pilus epitope (termed adhesintopes) using the PE and PCK portions of the *P. aeruginosa* pilin. These fusion proteins were selected as vaccine candidates in a mouse model. The PCK fusion proteins were compared to the previous adhesintope-based vaccine candidates against *P. aeruginosa*. The PCK fusion protein induced a 1000-fold greater antibody response, balanced in IgG1 and IgG2 antibody response. However, the immunized mice, when challenged with a *P. aeruginosa* infection, showed no significant level of protection (Bhatnagar, Awasthi, Nomellini, Smit, & Suresh, 2006; Umelo-Njaka *et al.*, 2001). These S-

layer proteins are also extensively studied in pathogen *Bacillus anthracis*, responsible for anthrax, a lethal infection of mammals. It has been noticed that the pathogen requires S-layer proteins for the pathogenesis and infection. Surface localization and abundant expression of the S-layer protein BslA make this polypeptide a candidate antigen for purified subunit vaccines against *B. anthracis* (Kern & Schneewind, 2008). Therefore, it is suggested that these proteins should be analysed experimentally. As supported by *in silico* analyses and track record in other organism, S-layer proteins are suitable candidates for *C. fetus* subspecies vaccine development.

Hsp12 (Heat shock protein 12) variant is also one of our identified candidates. However, the current literature lacks enough information about its role in bacterial survival and pathogenesis. But, it (heat shock protein 12.6; HSP12.6) has been used as vaccine antigen in combination with two well characterized vaccine antigens of *Brugia malayi* (causative agent of lymphatic filariasis), Abundant Larval transcript-2 (ALT-2) and tetraspanin large extra cellular loop (TSP-LEL) are used as multivalent fusion vaccine. Immunized individuals carry circulating antibodies against all three antigens. Hence, trivalent HAT vaccine is either used as a protein alone or as heterologous prime boost vaccine, conferring significant protection (95%) against *B. malayi* L3 challenge (Dakshinamoorthy, Samyikutty, Munirathinam, Reddy, & Kalyanasundaram, 2013).

Beside these prominent predicted candidate immunogenic genes/proteins, we suggest that other potential immunogenic sequences predicted in this study (<http://dx.doi.org/10.1016/j.gene.2012.07.070>.) should also be tested as they can serve as potential candidate for vaccines. We expect promising results from our candidates once they are subjected to experimental validations for eliciting immune responses in animal models. Our analysis of the exo-proteome of *C. fetus* subspecies has also led us to discover a number of core virulence factors, majority of which are even conserved across the genus.

Pathogenicity islands prediction in *C. fetus subspecies* was a good approach to identify

subspecies specific regions containing prominent virulence factors. By doing so, two potential PAIs (PICFV5 and PICFV8) in *C. fetus* subsp. *venerealis* and one (PICFF6) in *C. fetus* subsp. *fetus* were identified. We considered these islands due to the prominent gene contents. Beside these islands, a number of other predicted pathogenicity islands were observed not having direct relation to pathogenicity. Also, functions of the majority of genes of these islands are still unknown.

Pathogenicity island (PICFV5) was identified specifically in *C. fetus* subsp. *venerealis*. It was found to contain *sap* gene, which is partially deleted in *C. fetus* subsp. *fetus*. However, the region contains conserved *sap* gene in *C. fetus* subspecies. The *sap* gene encodes surface array proteins (SAP). This high molecular mass (97-149 kDa) surface array proteins (SAP) of *C. fetus* is critical to virulence (Kienesberger *et al.*, 2011). The observed capsule-like protein has role in pathogenesis and is responsible for immune evasion via antiphagocytic properties. It is long-known virulence factor and is found to be pathogenic to humans and animals (Fogg *et al.*, 1990). Neighboring genes to *sap* found in the same islands are annotated as hypothetical genes. It is suggested that these genes are functionally analysed in laboratory to check their immunogenicity.

Pathogenicity island (PICFV8) predicted in the pathogen *C. fetus* subsp. *venerealis* has a unique type secretion system (T4SS) which is absent in *C. fetus* subsp. *fetus*. The importance of Type IV secretion (T4SS) systems lies in its ability to transport DNAs and/or proteins through the membranes of bacteria (Gram-negative and Gram-positive bacteria). They form large multiprotein complexes consisting of 12 proteins termed VirB1-11 and VirD4. VirB7, 9 and 10 assembling into a 1.07 MegaDalton membrane-spanning core complex (CC), around which all other components assemble. Based on the function, the T4SS can be subdivided into (i) conjugation systems, (ii) DNA release or uptake systems and (iii) effector translocation systems. Conjugation systems mediate the transfer of DNA to recipient cells in a contact-dependent manner. These systems promote genome plasticity in bacteria and thus

mediate rapid adaptive responses to changes in environment. They are also responsible for the spread of antibiotic-resistance genes among pathogenic bacteria. T4SS systems, mediate DNA release and uptake and also contribute to the expansion of genome diversity (Kienesberger *et al.*, 2011; Rivera-Calzada *et al.*, 2013).

It is worthy to mention that *C. fetus* subsp. *venerealis* supports intra- and inter-species conjugative mobilization of plasmid DNA with the help of T4SS. Two putative secretion substrates encoded by the *C. fetus* subsp. *venerealis* PAI were investigated. The evidence for conjugative DNA delivery was found. However, no inter-bacterial protein transfer was observed (Kienesberger *et al.*, 2011).

The type IV secretions system is also characterized in human pathogen *H. pylori*. The *cag* island harbors genes for a type IV secretion system, mediating the translocation of the effector protein CagA (dark grey) into eukaryotic cells (Schmidt & Hensel, 2004).

The prominent island (PICFF6) predicted in *C. fetus* subsp. *fetus* contains Clustered Regularly Interspaced Short Palindrome Repeats (CRISPR) and a *dam* gene. The importance of this island is in the immune defense system of the bacteria. This is due to the fact that CRISPR loci and their associated *cas* (CRISPR associated) genes encode a sequence-specific defense mechanism against bacteriophage infection and plasmid conjugation (Bhaya & Barrangou, 2011; Nishida, 2012). CRISPR loci consists of arrays of short repetitive sequences (approximately 40 bp long) separated by equally short “spacer” sequences, many of them derived from mobile genetic elements such as bacteriophages and plasmids. However, the impact of CRISPRs on the emergence of virulence is not known yet. Recently, Bikard and colleagues conducted a study in the human pathogen *Streptococcus pneumoniae*. It was shown that CRISPR interference can prevent transformation of non-encapsulated, avirulent pneumococci into capsulated virulent strains during infection in mice. Additionally, it was also shown that at low frequencies bacteria can lose CRISPR function, acquire capsule genes and mount a successful infection. These results demonstrated that

CRISPR interference can prevent the emergence of virulence in vivo. Also, it was shown that strong selective pressure for virulence or antibiotic resistance can lead to loss of CRISPR in bacterial pathogens (Bikard et al., 2012). Similar studies could be effective in *C. fetus* subsp. *fetus* too.

The *dam* gene identified on the same island was found to have significant role in *C. fetus* subsp. *fetus* defense system by methylation of the DNA. The Dam enzyme (DNA methyltransferase) produced by *dam* gene catalyzes methylation of adenine residues in –GATC– sequences through the transfer of the methyl group from S-adenosyl-L-methionine (SAM), using substrate hemimethylated DNA (Soares et al., 2012). This specific methylation mechanism plays an important role in restriction–modification (R–M) systems. By this system, the foreign DNA is recognized (due to the lack of a different methylation pattern) and consequently cleaved (Kim et al., 2008).

In *E. coli* the *dam* gene encodes a DNA methyltransferase (DNA adenine methyltransferase) that methylates adenine in –GATC– sequences in double-stranded DNA mutant strains. Lack of this enzyme display a pleiotropic phenotype including increased mutability, hyper-recombination and increased sensitivity to DNA-damaging agents. Furthermore, *dam* bacteria have an increased number of single-strand breaks in DNA, compared to wild type. The regulation of the *dam* gene expression in *Escherichia coli* has been achieved with the *araBAD* promoter lacking a ribosome binding site. The *dam* is suggested to be targeted for laboratory mutagenesis to interfere with defense system of *C. fetus* subspecies (Marinus, 2000). It has been targeted for mutation in gene “cj1461”, a predicted methyltransferase gene reducing the motility of *C. jejuni* 81-176. Furthermore, electron microscopy revealed that the mutant strain has flagella but with aberrant structure. The Δ cj1461 mutant was sevenfold more adherent but 50-fold less invasive of INT-407 human epithelial cells than the wild type (Kim et al., 2008).

This study brought insights into the pathogenic life style of the *C. fetus* subspecies, together with common and unique virulence factors. Also, the mechanism of transport of molecule (DNA/protein) to target cells and the defense system machinery have been identified. In the light of current literature, it is possible to investigate the identified immunogenic targets for the development of proper therapeutic agents and eventually effective vaccines against these organisms.

Chapter 3 of this document highlighted the significance of human stomach colonizer *H. pylori*. This is an immediate threat to humans due the fact that these organism chronically infects the gastric mucosa in more than 50% of the human population (N. R. Salama et al., 2007). This strong association of the organism with its human host has suggested its coevolution through time. The *H. pylori* is responsible for 10% of all cancer-related deaths around the world (S. Zhang, Moise, & Moss, 2011). It is therefore important to identify *H. pylori* virulence factors, candidate genes and proteins for diagnostic and therapeutics purposes to get insight into the relatively unexplored genetic diversity of *H. pylori* (Lehours et al., 2011). Looking into the past record of developments in vaccine against this significant organism, relatively few efforts is observed. We therefore put our efforts to exploit the genomic sequences of multiple *H. pylori* strains and predicted essential genes and candidate sequences for therapeutic measures against this silent killer. Beside the target organism (*H. pylori*), other species from the genus *Helicobacter* are also included to understand relationships between them and with *H. pylori* species.

To the best of our knowledge, it was for the first time that someone applied pangenome reverse vaccinology approach to large number of *H. pylori* strains (38) to determine the core targets for development of antibiotics and vaccines (N. Salama et al., 2000).

The phylogenetic analysis demonstrated the positions of global representatives strains of *H. pylori* non-*pylori* species, based on sequences similarities in ribosomal RNA (Appendix) gene. Genomic variations (heatmap) revealed a significant diversity at genus level (Figure 1)

(N. R. Salama et al., 2007; Suerbaum & Josenhans, 2007). However, the interesting finding in this analysis is that all *H. pylori* strains distributed to two separate branches. A clear picture of genomic similarity is observed in the members based on our heatmap analysis. Example of these close phylogenetic relationships is observed in strains isolated from different regions in Peru (HP Cuz20, HP Shi 169, HP Shi 112, HP Shi 417, HP Shi 470 and HP Sat 464). However, the only strain isolated from Peru HP PeCan 18 was positioned on second branch away from other Peru strains, based on their genomic heterogeneity. It showed higher genomic similarities with West African strains (HP 2017, HP 2018, HP 908) and strain HP J99 isolated from patient in USA. Similar results were observed in a comparative analysis of HP 908 with HP J99, HP 908 (isolated from African patient living in France), revealing several specific genome features and novel insertion-deletion and substitution events. The genome sequence also revealed several strain-specific deletions and/or gain of genes exclusively present in HP908, compared with different sequenced genomes already available in the public domain (Devi *et al.*, 2010).

One of the objective of this analysis was to figure out the pattern of distribution of *H. pylori* isolated from different parts of the world in phylogenetic tree. However, in contrast to our expectation, the results could not be correlated with the geographical distribution of the organisms, except the strains isolated from Peru and West Africa. These strains showed close phylogenetic association based on the two applied approaches. A remarkable characteristic of *H. pylori* biology is its extraordinary allelic diversity and genetic variability. This diversity can be seen from the fact that almost every infected person harbour their own individual *H. pylori* strains (Hale *et al.*, 2012; Suerbaum & Josenhans, 2007). Another reason of this analysis was to visualize the genomic relatedness based on the genetic content similarities, which assisted us in further comparative analysis and correlations in results obtained in series of analysis.

Pangenome estimation of *H. pylori* genomes demonstrated that the organism (strains) shares (homology) much of their genomic contents and resulted in 1,185 conserved gene families. This could be seen clearly as core genome is ~77% of the average genome size and 42% of the species pangenome. This analysis indicated that individual genome carries ~33 percent of specific genes, which may result in diversity among strains. Based on these findings, according to our expectation, we can say that *H. pylori* strains contain much of the conserved information and the idea of estimating the pangenome was worthwhile. This conserved genomic data led us to the identification of candidate targets for development of broad spectrum drugs and vaccines development. Our results are somehow in agreement with the one predicted by Salama et al., in 2000. They determined the core set of genes in 15 *H. pylori* strains with a microarray method, and total of 1,281 genes were found to constitutes the core genome. Our predicted core genome strengthens their proposition that when more genome sequences become available the exact core genes can be estimated (N. Salama et al., 2000). The difference between the two analyses is almost 100 genes, which are decreased in the our core genome. We also observed that most of the genomes carry few new genes families in them, indicating genomic conservation in *H. pylori* intra-species and an average of 46 new gene families are added with subsequent genome. However, we are not confident to say that the same pattern will continue while more genomes are sequenced and analyzed in future. But we can somehow say that *H. pylori* genomes from different continents are diverse in nature, although at proteomic comparison revealed that they show higher proteomic similarities. For example, ~98% proteomic homology is observed in *H. pylori* 2017 and *H. pylori* 2018. Inter-species example of conservation is the *H. acinonychis* and *H. pylori* species, showing higher homologies based on heatmap data. A previous report showed that *H. acinonychis* shares 612 orthologues with *H. pylori*, and only few of their amino acids are different. This finding also supports recent possible host shifts from human to felines (Eppinger et al., 2006).

With the notion that core genome contains representative genes of all the organisms in the study and could be good target for broad spectrum vaccine and drugs targets, all completed genomes are examined to predict vaccine targets that are conserved (core) in all genomes (pangenome) (Donati & Rappuoli, 2013). For the first time, subtractive genomic approach is applied to *H. pylori* (genomes) and a central minimal genome composed of 493 gene families is calculated. The human genome sequence in the last decade provided greater opportunity for such analysis and hence potential non-host (human) homologs 250 genes, common to all *H. pylori* species, have been estimated (Barh, Tiwari, et al., 2011). The subcellular localization of conserved proteins led us to characterize the potential proteins suitable for drugs or vaccines development. Cytoplasmic proteins can act as possible drug targets, while surface membrane proteins can be used as vaccine targets. Therefore, our intention was to predict the best immunogenic proteins and vaccine candidates (Barinov *et al.*, 2009; He *et al.*, 2010). Proteins exposed on the surface play a role in helping to bypass the organism immune response or to facilitate infection. Therefore they are considered ideal targets for both immunological and therapeutic treatments. We used the a pipeline, SLEP (Surface Localization of External Protein) based on a combination of the best performing tools for microbial gene discovery, protein annotation and localization identification. The tool can be useful in prioritizing vaccine candidates as well as in identifying potential new targets for therapy. We were able to identify 88 surface proteins (Giombini *et al.*, 2010).

To predict the suitable candidates, the reverse vaccinology approach seems a suitable strategy and consequently it was applied with the following parameters: i) membrane, exported and lipo proteins, ii) transmembrane domains (not multiple helices), iii) adhesin quality, iv) conservation in all genomes, v) absence in non-pathogenic, vi) non-homologous to host (human), vii) MHC-I and II binding epitopes and viii) functional analysis (He *et al.*, 2010; Rinaudo *et al.*, 2009). The predicted potential target sequences with functional annotation are demonstrated (appendix IX.II) and can be considered as potential candidates

to develop effective drugs and vaccines.

Although *In silico*, these target sequences are predicted to be associated with vital cellular functions. However, detailed experimental evidence for their effectiveness is a subsequent stage. Interestingly, our predicted core target sequences contain proteins having major roles in transport systems. For example, type IV secretion system protein is an important factor associated with transport of proteins into eukaryotic (human) target cells. The type IV secretion system in *H. pylori* is capable of delivering the 128-kDa protein CagA into epithelial cells (Kutter *et al.*, 2008).

From pathogenomic perspective, most *H. pylori* strains that cause diseases contain the *cag* pathogenicity island (PAI), a chromosomal region comprising approximately 37,000 base pairs and 29 genes. A typical *cag* PAI encodes 27 genes, although islands with up to 33 genes have been reported. The significance of the *cag* genes is because of their involvement in the assembly of the type IV secretion system that translocates the protein CagA into the cytoplasm of gastric epithelial cells (Suerbaum and Josenhans 2007). At least six of genes on this island code for proteins with homology to type IV secretion system components. A systematic mutagenesis study has shown that 17 of the *cag* island genes are essential for translocation of CagA, and 14 are essential for the ability of *H. pylori* to strongly induce IL-8 secretion. *H. pylori* type IV secretion system has been observed on the surface of *cag* PAI-carrying strains (and were not observed in *cag* island deletion mutants). The *cag* PAI is genetically unstable due to 31-bp flanking repeats, which could be removed (RecA-dependent precise excision) (Gressmann *et al.*, 2005; You *et al.*, 2012). It is also shown that *H. pylori* presents all three types of T4SS system, although one of them is unique. Type I strains of the bacterium have a special T4SS called cytotoxin-associated gene-pathogenicity island (*cagPAI*). At least two different roles have been assigned to the *cagPAI*: it is used by the pathogen to inject a toxin into the host cell and it is also able to induce the production of interleukin-8 in humans (You *et al.*, 2012; Zanotti, 2011).

We attempted to identify more PAIs in *H. pylori* by using the same strategy which was applied to *Campylobacter fetus* subspecies. By doing so, total of 22 putative PAIs (PiHps) (appendix IX.III figure 3) were detected with reference to HP 26695 (appendix IX.III Figure 3). We showed that among the predicted PiHps, 2, 4, 14 and 15 are highly conserved among *H. pylori* stains and 8 and 13 are highly variable among 38 *H. pylori* strains analysed. The majority of the genes harbored by PiHps: 2, 4, 8, 13, 14 and 15 have been assigned the term “hypothetical protein” on their product tag, meaning that most of the gene products are not yet identified. Once these gene characterized, may give insights into the virulence factors in *H. pylori*. The conserved and variable islands containing genes might also helps in understanding of the selective (gastric and intestinal) behavior of the species (strains). it might also be possible that novel islands in *H. pylori* are discovered. This finding will definitely bring new insights into the *H. pylori* phylogenomics and pathogenic lifestyle, once fully characterized.

We wanted to conclude this session with general findings, observations and challenges while studying different genera (including multiple species):

- Phylogenomics plays an important role in comparative genomic analysis. It gives a general overview of the selected organisms (species) and helps in understanding the evolutionary relationships in presence of an outgroup. Moreover, during different analysis it can be revisited for consistency in the results. Based on our observations in all the three chapters, phylogenetic analyses played a major role in estimation of organism association through evolution. The analysis of the trees based on single gene (16S rRNA) is an easy and cheap method. Based on this, we observed that inter-species are at distant position while intra-species (strains) showed close relationships (Guindon *et al.*, 2010; Huson, 1998). The success in such comparative studies are based on the consistencies throughout the results, starting from single

gene (16s rRNA tree) analysis to that of core gene families and pan gene families (pangenome tree) distribution. Furthermore, the nature of the genome sequences (draft or finished) and multiple analysis to validate the results are influencing factors in obtaining optimum results. Multiple genome analyses are of much importance in providing the broader view of the gene pool of a species and hence the pangenome.

- Our observation and literature data are in agreement to reveal that the size of the core genome reflects the evolutionary history and lifestyle of each species. It can be as little as 42% of the genome in species such as *E. coli*, which is able to colonize in many different environments. If we consider the genera *Corynebacterium* and *Campylobacter*, the estimated core genomes (741 and 552 respectively) are ~31% of individual genome (2,391 and 1794 respectively), indicative of the genomic variability at the genus level (Donati & Rappuoli, 2013). Even more genomic variability was observed in *Helicobacter* species, the core genome is considerably low, as low as 29%. However, intra-species (*H. pylori*) core genome contains 77% genes from individual genome and only 33% of the genome varies (specific) in each genome. Although the core genome mostly encodes metabolic functions that are essential for cell viability, the rest of the genome which is defined as dispensable genome is comparatively rich in poorly characterized genes. Also, it contains genes associated with mobile and extra-chromosomal elements, supporting the hypothesis that the majority of strain-specific traits depend on lateral gene transfer events. Therefore, it is recommended to analyse the disposable genome between species to identify the character genes and genes responsible for the adaptation of bacteria to specific environments.
- From a vaccine discovery perspective, the core and pangenome can be seen both as a constraint and an opportunity. For example, a universal antigen, if required to predict, must be part of the conserved genome. Consequently, the panel of potential

candidates (pangenome) is restricted. In principle, vaccine candidates able to guarantee partial protection, can be derived from a gene pool that is larger than the genome of a single strain and alternative multicomponent vaccines including two or more of these antigens might guarantee broad protection. However, identification of the best combination requires a deep understanding of the epidemiology of the species, as an essential component of the initial screening of candidates (Donati & Rappuoli, 2013; He *et al.*, 2010; Rinaudo *et al.*, 2009; Santos *et al.*, 2011).

- One of the major challenges for such studies is to develop techniques for assessing the function of novel genes on large scale and integrating information on how genes and proteins interact at the cellular level to create and maintain a living organism.

In the end, the results described in this study aid our understanding of gene and protein content relationships in closely related bacterial groups, allowing us to make further inferences regarding genome-genome relationships, genome evolution, diversity among species, common and unique features and pathogenic lifestyles of distant but related bacterial species.

VI. CONCLUSION AND FUTURE PERSPECTIVES

This study illustrates the significance of comparative genomic strategies for studying diverse bacterial species and pangenomic analysis of closely related species. The applied tools and strategies discussed are relatively simple to implement for identification of genes in bacterial species and comparative analysis of multiple genomes. The results are reproducible to the best of our knowledge. However, they may support or contradict existing results, hypotheses and taxonomic divisions. They can also generate novel ideas and hypotheses.

In case of analysed veterinary pathogens, *Corynebacterium pseudotuberculosis*, *Campylobacter fetus* species and human pathogenic species *Helicobacter pylori* could be considered as models for analysis of other pathogenic species. Moreover, the data presented here can assist further in understanding the intra- and inter species relationships in bacterial populations and with their animal or human host. The strategies and methodologies applied in this study can also be applied to large numbers of genomes (pangenome analysis), while studying bacterial life style and general mechanisms to predict trends even across genera.

The pathogenomic strategies used here can also be extended to other species for the identification of common and unique genes and proteins linked to metabolism, defense mechanisms and pathogenicity. As an ongoing project from the laboratory (LGCM), the analysis completed to date will be extended to other *C. pseudotuberculosis* species which are in pipeline. However, we finalized this session with few suggestions and recommendations to the readers and to those who are interested in taking this research one step ahead.

Compared to *Corynebacterium* species, *Campylobacter fetus* subspecies are less characterized. Therefore, we suggest more complete genome sequences of *C. fetus* subspecies for deep insights into the subspecies pangenome. Unique genes identified in this study could be functionally characterized to understand the mechanism of host specificity in subspecies.

In case of *H. pylori*, the strains are of diverse nature and distribution. However, the higher genomic similarities and relatively greater conserved genome could be further exploited to obtain potential information regarding common virulence factors, diagnostics, drugs and vaccines. Furthermore, the pangenomic data can assist in studies to examine the geographical distribution and diversity in *H. pylori* species.

There is also need of a multi-step and robust protocol to analyse the pangenome in extended bacterial population. Moreover, an immediate task for the bioinformatics researcher is to integrate all the mentioned programs (steps) in this project into a single efficient pipeline, in order to minimize the labor and save the time.

VII. BIBLIOGRAPHY

- Abby, S., & Daubin, V. (2007). Comparative genomics and the evolution of prokaryotes. *Trends in microbiology*, 15(3), 135–41. doi:10.1016/j.tim.2007.01.007
- Acencio, M. L., & Lemke, N. (2009a). Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. *BMC bioinformatics*, 10, 290. doi:10.1186/1471-2105-10-290
- Acencio, M. L., & Lemke, N. (2009b). Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. *BMC Bioinformatics*, 10(1), 290.
- Agren, J., Sundström, A., Håfström, T., & Segerman, B. (2012). Gegenees: fragmented alignment of multiple genomes for determining phylogenomic distances and genetic signatures unique for specified target groups. *PloS one*, 7(6), e39107. doi:10.1371/journal.pone.0039107
- Ali, A., Soares, S. C., Barbosa, E., Santos, A. R., Barh, D., Bakhtiar, S. M., Hassan, S. S., et al. (2013). Bacteriology & Parasitology Microbial Comparative Genomics : An Overview of Tools and Insights Into The Genus *Corynebacterium*, 4, 1–16. doi:10.4172/2155-9597.1000167
- Ali, A., Soares, S. C., Santos, A. R., Guimarães, L. C., Barbosa, E., Almeida, S. S., Abreu, V. A. C., et al. (2012). *Campylobacter fetus* subspecies: Comparative genomics and prediction of potential virulence targets. *Gene*. doi:10.1016/j.gene.2012.07.070
- Arnold, D. L., & Jackson, R. W. (2011). Bacterial genomes: evolution of pathogenicity. *Current Opinion in Plant Biology*, 14(4), 385–391.
- Azevedo, V., Abreu, V., Almeida, S., Santos, A., Soares, S., Ali, A., Pinto, A., et al. (2011). Whole Genome Annotation: In Silico Analysis.
- Barcellos, F. G., Menna, P., Da Silva Batista, J. S., & Hungria, M. (2007). Evidence of horizontal transfer of symbiotic genes from a *Bradyrhizobium japonicum* inoculant strain to indigenous diazotrophs *Sinorhizobium* (Ensifer) *fredii* and *Bradyrhizobium elkanii* in a Brazilian Savannah soil. *Applied and environmental microbiology*, 73(8), 2635–43. doi:10.1128/AEM.01823-06
- Barh, D., Jain, N., Tiwari, S., Parida, B. P., D'Afonseca, V., Li, L., Ali, A., et al. (2011). A novel comparative genomics analysis for common drug and vaccine targets in *Corynebacterium pseudotuberculosis* and other CMN group of human pathogens. *Chemical biology & drug design*.

- Barh, D., Tiwari, S., Jain, N., Ali, A., Santos, A. R., Misra, A. N., Azevedo, V., et al. (2011). In silico subtractive genomics for target identification in human bacterial pathogens. *Drug Development Research*, 72(2), 162–177.
- Barinov, A., Loux, V., Hammani, A., Nicolas, P., Langella, P., Ehrlich, D., Maguin, E., et al. (2009). Prediction of surface exposed proteins in *Streptococcus pyogenes*, with a potential application to other Gram-positive bacteria. *Proteomics*, 9(1), 61–73. doi:10.1002/pmic.200800195
- Bhatnagar, P. K., Awasthi, A., Nomellini, J. F., Smit, J., & Suresh, M. R. (2006). Anti-tumor effects of the bacterium *Caulobacter crescentus* in murine tumor models. *Cancer biology & therapy*, 5(5), 485–91.
- Bhaya, D., Davison, M., & Barrangou, R. (2011). CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. *Annual review of genetics*, 45, 273–97. doi:10.1146/annurev-genet-110410-132430
- Bikard, D., Hatoum-Aslan, A., Mucida, D., & Marraffini, L. A. (2012). CRISPR interference can prevent natural transformation and virulence acquisition during in vivo bacterial infection. *Cell host & microbe*, 12(2), 177–86. doi:10.1016/j.chom.2012.06.003
- Blaser, M. J., Wang, E., Tummuru, M. K., Washburn, R., Fujimoto, S., & Labigne, A. (1994). High-frequency S-layer protein variation in *Campylobacter fetus* revealed by *sapA* mutagenesis. *Molecular microbiology*, 14(3), 453–62.
- Buck, G. A., Cross, R. E., Wong, T. P., Loera, J., & Groman, N. (1985). DNA relationships among some tox-bearing corynebacteriophages. *Infection and Immunity*, 49(3), 679–684.
- Chakrabarty, A. M. (1998). Nucleoside diphosphate kinase: role in bacterial growth, virulence, cell signalling and polysaccharide synthesis. *Molecular microbiology*, 28(5), 875–82.
- Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y., & Jin, Q. (2005). VFDB: a reference database for bacterial virulence factors. *Nucleic acids research*, 33(Database issue), D325–8. doi:10.1093/nar/gki008
- Chen, S. L., Hung, C.-S., Xu, J., Reigstad, C. S., Magrini, V., Sabo, A., Blasiar, D., et al. (2006). Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: a comparative genomics approach. *Proceedings of the National*

- Academy of Sciences of the United States of America*, 103(15), 5977–82. doi:10.1073/pnas.0600938103
- D'Auria, G., Jiménez-Hernández, N., Peris-Bondia, F., Moya, A., & Latorre, A. (2010). *Legionella pneumophila* pangenome reveals strain-specific virulence factors. *BMC genomics*, 11, 181. doi:10.1186/1471-2164-11-181
- Dakshinamoorthy, G., Samykutty, A. K., Munirathinam, G., Reddy, M. V., & Kalyanasundaram, R. (2013). Multivalent fusion protein vaccine for lymphatic filariasis. *Vaccine*, 31(12), 1616–22. doi:10.1016/j.vaccine.2012.09.055
- Deng, W., Puente, J. L., Gruenheid, S., Li, Y., Vallance, B. a, Vázquez, A., Barba, J., et al. (2004). Dissecting virulence: systematic and functional analyses of a pathogenicity island. *Proceedings of the National Academy of Sciences of the United States of America*, 101(10), 3597–602. doi:10.1073/pnas.0400326101
- Devi, S. H., Taylor, T. D., Avasthi, T. S., Kondo, S., Suzuki, Y., Megraud, F., & Ahmed, N. (2010). Genome of *Helicobacter pylori* strain 908. *Journal of bacteriology*, 192(24), 6488–9. doi:10.1128/JB.01110-10
- Dobrindt, U., Janke, B., Piechaczek, K., Nagy, G., Ziebuhr, W., Fischer, G., Schierhorn, A., et al. (2000). Toxin genes on pathogenicity islands: impact for microbial evolution. *International journal of medical microbiology : IJMM*, 290(4-5), 307–11.
- Donati, C., & Rappuoli, R. (2013a). Reverse vaccinology in the 21st century: improvements over the original design. *Annals of the New York Academy of Sciences*, 1–18. doi:10.1111/nyas.12046
- Donati, C., & Rappuoli, R. (2013b). Reverse vaccinology in the 21st century : improvements over the original design, 1–18. doi:10.1111/nyas.12046
- Dong, Q.-J. (2009). Comparative genomics of *Helicobacter pylori*. *World Journal of Gastroenterology*, 15(32), 3984. doi:10.3748/wjg.15.3984
- Dworkin, J., Shedd, O. L., & Blaser, M. J. (1997). Nested DNA inversion of *Campylobacter fetus* S-layer genes is recA dependent. *Journal of bacteriology*, 179(23), 7523–9.
- Edgar, R. C., & Sjölander, K. (2004). COACH: profile-profile alignment of protein families using hidden Markov models. *Bioinformatics*, 20(8), 1309–18. doi:10.1093/bioinformatics/bth091

- Eppinger, M., Baar, C., Linz, B., Raddatz, G., Lanz, C., Keller, H., Morelli, G., et al. (2006). Who ate whom? Adaptive *Helicobacter* genomic changes that accompanied a host jump from early humans to large felines. *PLoS genetics*, 2(7), e120. doi:10.1371/journal.pgen.0020120.
- Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science (New York, N.Y.)*, 269(5223), 496–512.
- Fogg, G. C., Yang, L. Y., Wang, E., & Blaser, M. J. (1990). Surface array proteins of *Campylobacter fetus* block lectin-mediated binding to type A lipopolysaccharide. *Infection and immunity*, 58(9), 2738–44.
- Forde, B. M., & O'Toole, P. W. (2013). Next-generation sequencing technologies and their impact on microbial genomics. *Briefings in functional genomics*. doi:10.1093/bfpg/els062
- Fraser, C. M., Eisen, J., Fleischmann, R. D., Ketchum, K. A., & Peterson, S. (2000). Comparative genomics and understanding of microbial biology. *Emerging Infectious Diseases*, 6(5), 505–512.
- Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., et al. (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science (New York, N.Y.)*, 270(5235), 397–403.
- Friis, C., Jensen, L. J., & Ussery, D. W. (2000). Visualization of pathogenicity regions in bacteria. *Genetica*, 108(1), 47–51.
- Gardy, J. L., Laird, M. R., Chen, F., Rey, S., Walsh, C. J., Ester, M., & Brinkman, F. S. L. (2005). PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics (Oxford, England)*, 21(5), 617–23. doi:10.1093/bioinformatics/bti057
- Giombini, E., Orsini, M., Carrabino, D., & Tramontano, A. (2010). An automatic method for identifying surface proteins in bacteria: SLEP. *BMC bioinformatics*, 11, 39. doi:10.1186/1471-2105-11-39
- Gorkiewicz, G., Kienesberger, S., Schober, C., Scheicher, S. R., Güllý, C., Zechner, R., & Zechner, E. L. (2010). A genomic island defines subspecies-specific virulence features of the host-adapted pathogen *Campylobacter fetus* subsp. *venerealis*. *Journal of bacteriology*, 192(2), 502–17. doi:10.1128/JB.00803-09

- Grant, J. R., Arantes, A. S., & Stothard, P. (2012). Comparing thousands of circular genomes using the CGView Comparison Tool. *BMC genomics*, 13(1), 202. doi:10.1186/1471-2164-13-202
- Gressmann, H., Linz, B., Ghai, R., Pleissner, K.-P., Schlapbach, R., Yamaoka, Y., Kraft, C., et al. (2005). Gain and loss of multiple genes during the evolution of *Helicobacter pylori*. *PLoS genetics*, 1(4), e43. doi:10.1371/journal.pgen.0010043
- Groman, N., Schiller, J., & Russell, J. (1984). *Corynebacterium ulcerans* and *Corynebacterium pseudotuberculosis* responses to DNA probes derived from corynephage beta and *Corynebacterium diphtheriae*. *Infection and immunity*, 45(2), 511–7.
- Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology*, 59(3), 307–21. doi:10.1093/sysbio/syq010
- Haag, N. L., Velk, K. K., & Wu, C. (2011). In silico Identification of Drug Targets in Methicillin / Multidrug-Resistant *Staphylococcus aureus*, (c), 91–99.
- Hacker, J., & Carniel, E. (2001). Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. *EMBO reports*, 2(5), 376–81. doi:10.1093/embo-reports/kve097
- Hale, J., Suerbaum, S., Mugisha, L., Schlebusch, C. M., Bernho, S., Merwe, S. W. Van Der, & Achtman, M. (2012). Age of the Association between *Helicobacter pylori* and, 8(5). doi:10.1371/journal.ppat.1002693
- He, Y., Xiang, Z., & Mobley, H. L. T. (2010). Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development. *Journal of biomedicine & biotechnology*, 2010, 297505. doi:10.1155/2010/297505
- Hughes, G. M., Gang, L., Murphy, W. J., Higgins, D. G., & Teeling, E. C. (2013). Using Illumina next generation sequencing technologies to sequence multigene families in de novo species. *Molecular Ecology Resources*, n/a–n/a. doi:10.1111/1755-0998.12087
- Huson, D. H. (1998). SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics (Oxford, England)*, 14(1), 68–73.

- Hyatt, D., LoCascio, P. F., Hauser, L. J., & Uberbacher, E. C. (2012). Gene and Translation Initiation Site Prediction in Metagenomic Sequences. *Bioinformatics*, 28(17), 2223–2230. doi:10.1093/bioinformatics/bts429
- Jackson, R. W., Johnson, L. J., Clarke, S. R., & Arnold, D. L. (2011). Bacterial pathogen evolution: breaking news. *Trends in genetics: TIG*, 27(1), 32–40. doi:10.1016/j.tig.2010.10.001
- Jensen, L. J., Friis, C., & Ussery, D. W. (1999). Three views of microbial genomes. *Research in Microbiology*, 150(9-10), 773–777.
- Jolley, K. A., Bliss, C. M., Bennett, J. S., Bratcher, H. B., Brehony, C., Colles, F. M., Wimalaratna, H., et al. (2012). Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology (Reading, England)*, 158(Pt 4), 1005–15. doi:10.1099/mic.0.055459-0
- Kawai, M., Furuta, Y., Yahara, K., Tsuru, T., Oshima, K., Handa, N., Takahashi, N., et al. (2011). Evolution in an oncogenic bacterial species with extreme genome plasticity: *Helicobacter pylori* East Asian genomes. *BMC microbiology*, 11(1), 104. doi:10.1186/1471-2180-11-104
- Kern, J. W., & Schneewind, O. (2008). BslA, a pXO1-encoded adhesin of *Bacillus anthracis*. *Molecular microbiology*, 68(2), 504–15. doi:10.1111/j.1365-2958.2008.06169.x
- Kienesberger, S., Schober Trummer, C., Fauster, A., Lang, S., Sprenger, H., Gorkiewicz, G., & Zechner, E. L. (2011). Interbacterial macromolecular transfer by the *Campylobacter fetus* subsp. *venerealis* type IV secretion system. *Journal Of Bacteriology*, 193(3), 744–758.
- Kim, J.-S., Li, J., Barnes, I. H. A., Baltzegar, D. A., Pajaniappan, M., Cullen, T. W., Trent, M. S., et al. (2008). Role of the *Campylobacter jejuni* Cj1461 DNA methyltransferase in regulating virulence characteristics. *Journal of bacteriology*, 190(19), 6524–9. doi:10.1128/JB.00765-08
- Koonin, E. V., & Wolf, Y. I. (2008). Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic acids research*, 36(21), 6688–719. doi:10.1093/nar/gkn668
- Krizova, L., & Nemec, A. (2010). A 63 kb genomic resistance island found in a multidrug-resistant *Acinetobacter baumannii* isolate of European clone I from 1977. *The Journal of antimicrobial chemotherapy*, 65(9), 1915–8. doi:10.1093/jac/dkq223

- Kuska, B. (1998). Beer, Bethesda, and biology: how “genomics” came into being. *Journal of the National Cancer Institute*, 90(2), 93.
- Kutter, S., Buhrdorf, R., Haas, J., Schneider-Brachert, W., Haas, R., & Fischer, W. (2008). Protein subassemblies of the *Helicobacter pylori* Cag type IV secretion system revealed by localization and interaction studies. *Journal of bacteriology*, 190(6), 2161–71. doi:10.1128/JB.01341-07
- Lagesen, K., Hallin, P., Rødland, E. A., Staerfeldt, H.-H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic acids research*, 35(9), 3100–8. doi:10.1093/nar/gkm160
- Lapierre, P., & Gogarten, J. P. (2009). Estimating the size of the bacterial pan-genome. *Trends in Genetics*, 25(3), 107–110.
- Lehours, P., Vale, F. F., Bjursell, M. K., Melefors, O., Advani, R., Glavas, S., & Guegueniat, J. (2011). Genome Sequencing Reveals a Phage in *Helicobacter pylori*, 2(6), 1–11. doi:10.1128/mBio.00239-11.Editor
- Lu, G., Jiang, L., Helikar, R. M. K., Rowley, T. W., Zhang, L., Chen, X., & Moriyama, E. N. (2006). GenomeBlast: a web tool for small genome comparison. *BMC bioinformatics*, 7 Suppl 4, S18. doi:10.1186/1471-2105-7-S4-S18
- Marinus, M. G. (2000). Recombination is essential for viability of an *Escherichia coli* dam (DNA adenine methyltransferase) mutant. *Journal of bacteriology*, 182(2), 463–8.
- Maurelli, A. T. (1998). “Black holes” and bacterial pathogenicity: A large genomic deletion that enhances the virulence of *Shigella* spp. and enteroinvasive *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 95(7), 3943–3948. doi:10.1073/pnas.95.7.3943
- McKusick, V. A., & Ruddle, F. H. (1987). Toward a complete map of the human genome. *Genomics*, 1(2), 103–6.
- Metzker, M. L. (2005). Emerging technologies in DNA sequencing. *Genome Research*, 15(12), 1767–1776.
- Moolhuijzen, P. M., Lew-Tabor, A. E., Wlodek, B. M., Agüero, F. G., Comerci, D. J., Ugalde, R. A., Sanchez, D. O., et al. (2009). Genomic analysis of *Campylobacter fetus* subspecies: identification of candidate virulence determinants and diagnostic assay targets. *BMC Microbiology*, 9, 86.

- Mushegian, a R., & Koonin, E. V. (1996). A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 93(19), 10268–73.
- Muzzi, A., Masignani, V., & Rappuoli, R. (2007). The pan-genome: towards a knowledge-based discovery of novel targets for vaccines and antibacterials. *Drug discovery today*, 12(11-12), 429–39. doi:10.1016/j.drudis.2007.04.008
- Narayanan, R., & Ramaswami, M. (2003). Regulation of dynamin by nucleoside diphosphate kinase. *Journal of bioenergetics and biomembranes*, 35(1), 49–55.
- Nishida, H. (2012). Genome DNA Sequence Variation, Evolution, and Function in Bacteria and Archaea. *Current issues in molecular biology*, 15(1), 19–24.
- On, S. L. W. (2001). Taxonomy of Campylobacter, Arcobacter, Helicobacter and related bacteria: current status, future prospects and immediate concerns. *Journal of Applied Microbiology*, 90(S6), 1S–15S. doi:10.1046/j.1365-2672.2001.01349.x
- Ott, L., McKenzie, A., Baltazar, M. T., Britting, S., Bischof, A., Burkovski, A., & Hoskisson, P. A. (2012). Evaluation of invertebrate infection models for pathogenic corynebacteria. *FEMS immunology and medical microbiology*, 65(3), 413–21. doi:10.1111/j.1574-695X.2012.00963.x
- Pagani, I., Liolios, K., Jansson, J., Chen, I.-M. A., Smirnova, T., Nosrat, B., Markowitz, V. M., et al. (2012). The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic acids research*, 40(Database issue), D571–9. doi:10.1093/nar/gkr1100
- Pallen, M. J., & Wren, B. W. (2007). Bacterial pathogenomics. *Nature*, 449(7164), 835–842.
- Raskin, D. M., Seshadri, R., Pukatzki, S. U., & Mekalanos, J. J. (2006). Review Bacterial Genomics and Pathogen Evolution. *Cell*, 124(4), 703–714. doi:10.1016/j.cell.2006.02.002
- Rinaudo, C. D., Telford, J. L., Rappuoli, R., & Seib, K. L. (2009). Vaccinology in the genome era. *The Journal of clinical investigation*, 119(9), 2515–25. doi:10.1172/JCI38330
- Rivera-Calzada, A., Fronzes, R., Savva, C. G., Chandran, V., Lian, P. W., Laeremans, T., Pardon, E., et al. (2013). Structure of a bacterial type IV secretion core complex at subnanometre resolution. *The EMBO journal*, 1–10. doi:10.1038/emboj.2013.58

- Röttger, R., Kalaghatgi, P., Sun, P., Soares, S. de C., Azevedo, V., Wittkop, T., & Baumbach, J. (2013). Density parameter estimation for finding clusters of homologous proteins--tracing actinobacterial pathogenicity lifestyles. *Bioinformatics (Oxford, England)*, 29(2), 215–22. doi:10.1093/bioinformatics/bts653
- Rottger, R., Ruckert, U., Taubert, J., & Baumbach, J. (2012). How Little Do We Actually Know? -- On the Size of Gene Regulatory Networks. *IEEEACM transactions on computational biology and bioinformatics IEEE ACM*, 9(5), 1293–1300. doi:10.1109/TCBB.2012.71
- Ruiz, J. C., D'Afonseca, V., Silva, A., Ali, A., Pinto, A. C., & others. (2011). Evidence for Reductive Genome Evolution and Lateral Acquisition of Virulence Functions in.
- Salama, N., Guillemin, K., McDaniel, T. K., Sherlock, G., Tompkins, L., & Falkow, S. (2000). A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. *Proceedings of the National Academy of Sciences of the United States of America*, 97(26), 14668–73. doi:10.1073/pnas.97.26.14668
- Salama, N. R., Gonzalez-Valencia, G., Deatherage, B., Aviles-Jimenez, F., Atherton, J. C., Graham, D. Y., & Torres, J. (2007). Genetic analysis of *Helicobacter pylori* strain populations colonizing the stomach at different times postinfection. *Journal of bacteriology*, 189(10), 3834–45. doi:10.1128/JB.01696-06
- Santos, A., Ali, A., Barbosa, E., Silva, A., Miyoshi, A., Barh, D., & Azevedo, V. (2011). THE REVERSE VACCINOLOGY--A CONTEXTUAL OVERVIEW.
- Schmidt, H., & Hensel, M. (2004a). Pathogenicity Islands in Bacterial Pathogenesis, 17(1), 14–56. doi:10.1128/CMR.17.1.14
- Schmidt, H., & Hensel, M. (2004b). Pathogenicity Islands in Bacterial Pathogenesis. *Society*, 17(1), 14–56. doi:10.1128/CMR.17.1.14
- Skirrow, M. B. (1994). Diseases due to *Campylobacter*, *Helicobacter* and related bacteria. *Journal of comparative pathology*, 111(2), 113–49.
- Snipen, L., & Ussery, D. W. (2010). Standard operating procedure for computing pangenome trees. *Standards in genomic sciences*, 2(1), 135–141.
- Soares, S. C., Silva, A., Trost, E., Blom, J., Ramos, R., Carneiro, A., Ali, A., et al. (2013). The Pan-Genome of the Animal Pathogen *Corynebacterium pseudotuberculosis* Reveals

- Differences in Genome Plasticity between the Biovar ovis and equi Strains. *PloS one*, 8(1), e53818. doi:10.1371/journal.pone.0053818
- Stynen, A. P. R., Lage, A. P., Moore, R. J., Rezende, A. M., De Resende, V. D. D. S., Ruy, P. D. C., Daher, N., et al. (2011). Complete genome sequence of type strain *Campylobacter fetus* subsp. *venerealis* NCTC 10354T. *Journal of bacteriology*, 193(20), 5871–2. doi:10.1128/JB.05854-11
- Suerbaum, S., & Josenhans, C. (2007). *Helicobacter pylori* evolution and phenotypic diversification in a changing host. *Nature reviews. Microbiology*, 5(6), 441–52. doi:10.1038/nrmicro1658
- Sun, J., Wang, X., Lau, A., Liao, T.-Y. A., Bucci, C., & Hmama, Z. (2010). Mycobacterial nucleoside diphosphate kinase blocks phagosome maturation in murine RAW 264.7 macrophages. *PloS one*, 5(1), e8769. doi:10.1371/journal.pone.0008769
- Tang, H., Lyons, E., Pedersen, B., Schnable, J. C., Paterson, A. H., & Freeling, M. (2011). Screening synteny blocks in pairwise genome comparisons through integer programming. *BMC bioinformatics*, 12(1), 102. doi:10.1186/1471-2105-12-102
- Terashima, H., Kojima, S., & Homma, M. (2008). Flagellar motility in bacteria structure and function of flagellar motor. *International review of cell and molecular biology*, 270, 39–85. doi:10.1016/S1937-6448(08)01402-0
- Tettelin, H., Massignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., et al. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome”. *Proceedings of the National Academy of Sciences of the United States of America*, 102(39), 13950–5. doi:10.1073/pnas.0506758102
- The Sanger Centre, W. (1998). Toward a complete human genome sequence. *Genome Research*, 8(11), 1097–1108.
- Trost, B., Haakensen, M., Pittet, V., Ziola, B., & Kusalik, A. (2010). Analysis and comparison of the pan-genomic properties of sixteen well-characterized bacterial genera. *BMC Microbiology*, 10(1), 258.
- Trost, E., Blom, J., Soares, S. de C., Huang, I.-H., Al-Dilaimi, A., Schröder, J., Jaenicke, S., et al. (2012). Pangenomic study of *Corynebacterium diphtheriae* that provides insights into the genomic diversity of pathogenic isolates from cases of classical diphtheria,

- endocarditis, and pneumonia. *Journal of bacteriology*, 194(12), 3199–215. doi:10.1128/JB.00183-12
- Tu, Z.-C., Gaudreau, C., & Blaser, M. J. (2005). Mechanisms underlying *Campylobacter* fetus pathogenesis in humans: surface-layer protein variation in relapsing infections. *The Journal of infectious diseases*, 191(12), 2082–9. doi:10.1086/430349
- Umelo-Njaka, E., Nomellini, J. F., Bingle, W. H., Glasier, L. G., Irvin, R. T., & Smit, J. (2001). Expression and testing of *Pseudomonas aeruginosa* vaccine candidate proteins prepared with the *Caulobacter crescentus* S-layer protein expression system. *Vaccine*, 19(11-12), 1406–15.
- Urwin, R., & Maiden, M. C. J. (2003). Multi-locus sequence typing: a tool for global epidemiology. *Trends in microbiology*, 11(10), 479–87.
- Ussery, D. W., Wassenaar, T. M., & Borini, S. (2008). *Computing for Comparative Microbial Genomics: Bioinformatics for Microbiologists (Computational Biology)* (p. 288). Springer.
- Wagner, K. S., White, J. M., Crowcroft, N. S., De Martin, S., Mann, G., & Efstratiou, A. (2010). Diphtheria in the United Kingdom, 1986-2008: the increasing role of *Corynebacterium ulcerans*. *Epidemiology and infection*, 138(11), 1519–30. doi:10.1017/S0950268810001895
- Watson, J. D., and Crick, F. H. C. A structure for deoxyribose nucleic acid. 1953. *Nature* 171:173
- Wen, Z., Wang, K., Li, M., Nie, F., & Yang, Y. (2005). Analyzing functional similarity of protein sequences with discrete wavelet transform. *Computational Biology and Chemistry*, 29(3), 220–228.
- Yadav, S. P. (2007). The wholeness in suffix -omics, -omes, and the word om. *Journal of biomolecular techniques : JBT*, 18(5), 277.
- You, Y., He, L., Zhang, M., Fu, J., Gu, Y., Zhang, B., Tao, X., et al. (2012). Comparative Genomics of *Helicobacter pylori* Strains of China Associated with Different Clinical Outcome. *PloS one*, 7(6), e38528. doi:10.1371/journal.pone.0038528
- Zanotti, G. (2011). Molecular aspects of *Helicobacter pylori* cag-pathogenicity island. *The FEBS journal*, 278(8), 1189. doi:10.1111/j.1742-4658.2011.08034.x

- Zhang, R., & Lin, Y. (2009). DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic Acids Research*, 37(Database issue), D455–D458.
- Zhang, S., Moise, L., & Moss, S. F. (2011). H. pylori vaccines: why we still don't have any. *Human vaccines*, 7(11), 1153–7. doi:10.4161/hv.7.11.17655
- Zheng, L.-L., Li, Y.-X., Ding, J., Guo, X.-K., Feng, K.-Y., Wang, Y.-J., Hu, L.-L., et al. (2012). A comparison of computational methods for identifying virulence factors. *PloS one*, 7(8), e42517. doi:10.1371/journal.pone.0042517
- Zhou, C. E., Smith, J., Lam, M., Zemla, A., Dyer, M. D., & Slezak, T. (2007). MvirDB--a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic acids research*, 35(Database issue), D391–4. doi:10.1093/nar/gkl791

VIII. CURRICULUM VITAE

Amjad Ali

Ali Amjad; Ali, A

Laboratory of Molecular and Cellular Genetics (LGCM)

Department of General Biology, ICB/UFMG,

Av. Antonio Carlos, 6627. Pampulha,

Belo Horizonte, Minas Gerais, Brazil.

CP 486 CEP 31270-901

Cell # +55 31 97564437

Date of Birth: June 1, 1982

Nationality: Pakistani

E-mail: amjad_uni@yahoo.com

Educational Background

PhD. Genetics/Genomics, Federal University of Minas Gerais, Brazil 2010-2013

MPhil. Biotechnology/Genetics, Quaid-i-Azam University, Islamabad 2006-2008

BS (Hons.) Biotechnology, University of Malakand, KPK 2001-2006

Professional Experience

PhD Research fellow, Federal University of Minas Gerais, BH, Minas Gerais, Brazil (2010-2013)

Senior Research Associate, Human Molecular Genetics Laboratory, NIBGE, Faisalabad, (2009)

Research Associate, HMG Laboratory, Health Biotechnology Division, NIBGE, Faisalabad. (2008)

Technical Support Executive, BIO-RAD Inc.; USA, M/S Chemical House Lahore, Pakistan (2007)

Research Interests: Genetics and Genomics

Genetics, Genomics and Microbial Genomics: Genome sequencing, Assembly and Annotation. Pan-genomics and Comparative Genomics. Pathogenomics of Bacterial Species.

Publications:

1. **Amjad Ali**, Siomar C Soares, Eudes Barbosa, Anderson R Santos, Debmalya Barh³, Syeda M. Bakhtiar¹, Syed S. Hassan, David W. Ussery, Artur Silva, Anderson Miyoshi, Vasco Azevedo, **Microbial Comparative Genomics: An Overview of Tools and Insights into the Genus *Corynebacterium***. J Bacteriol Parasitol 2013, 4;167. doi: 10.4172/2155-9597.1000167.
2. Bakhtiar SM, Leblanc JG, Salvucci E, **Ali Amjad**, Martin R, Langella P, Chatel JM, Miyoshi A, Bermúdez-Humarán LG, Azevedo V. **Implications of the human microbiome in Inflammatory Bowel Diseases**. FEMS Microbiol Lett. 2013 Feb 23. doi: 10.1111/1574-6968.12111
3. Soares SC, Silva A, Trost E, Blom J, Ramos R, Carneiro A, **Amjad Ali**, Santos AR, Pinto AC, Diniz C, Barbosa EG, Dorella FA, Aburjaile F, Rocha FS, Nascimento KK, Guimarães LC, Almeida S, Hassan SS, Bakhtiar SM, Pereira UP, Abreu VA, Schneider MP, Miyoshi A, Tauch A, Azevedo V. **The Pan-Genome of the Animal Pathogen *Corynebacterium pseudotuberculosis* Reveals Differences in Genome Plasticity between the Biovar ovis and equi Strains**. PLoS One. 2013;8(1):e53818. doi: 10.1371/journal.pone.0053818
4. **Amjad Ali**, Siomar C Soares, Syeda M Bakhtiar, Syed S. Hassan, Ulisses pereira, David W Ussery, Debmalya Barh, Artur Silva, Anderson Miyoshi, Vasco Azevedo, **Computational comparative genomic based insights into human gastric pathogen *Helicobacter pylori* (38 species), essential features and species pangenome. (Manuscript. 2013)**
5. **Amjad Ali**, Siomar C Soares, Anderson R Santos, Luis C Guimarães, Eudes Barbosa¹, Sintia S Almeida, Vinícius AC Abreu¹, Adriana, R Carneiro, Rommel TJ Ramos, Syeda M Bakhtiar, Syed S Hassan, David W Ussery, Stephen On⁴, Artur Silva², Maria P Schneider, Andrey P Lage, Anderson Miyoshi¹, Vasco Azevedo. **Campylobacter fetus subspecies: Comparative Genomics and Prediction of Potential Virulence Targets**. Gene 2012.
6. Barh D, Gupta K, Jain N, Khatri G, León-Sicairens N, Canizalez-Roman A, Tiwari S,

- Verma A, Rahangdale S, Shah Hassan S, Rodrigues Dos Santos A, **Amjad Ali**, Carlos Guimarães L, Thiago Jucá Ramos R, Devarapalli P, Barve N, Bakhtiar SM, Kumavath R, Ghosh P, Miyoshi A, Silva A, Kumar A, Narayan Misra A, Blum K, Baumbach J, Azevedo V. **Conserved host-pathogen PPIs**. Integr Biol (Camb). 2013.
7. Hassan SS, Guimarães Luis, Pereira Ulisses, Islam A, **Amjad Ali**, Syeda Marriam Bakhtiar, Dayana Ribeiro, Anderson Santos, Siomar Soares, Fernanda Dorella, Anne Pinto, Maria Schneider, Maria Barbosa, SÍntia Almeida, Vinícius Abreu, Flávia, Debmalya Barh, Anderson Barh, Anderson Miyoshi, Borna Miller, Artur Silva, Vasco Azevedo. **Complete genome sequence of *Corynebacterium pseudotuberculosis* biovar ovis strain P54B96 isolated from antelope in South Africa obtained by Rapid Next Generation Sequencing Technology**. SEGS 2012. doi:10.4056/sigs.3066455
 8. Soares SC, Trost E, Ramos RT, Carneiro AR, Santos AR, Pinto AC, Barbosa E, Aburjaile F, **Amjad Ali**, Diniz CA, Hassan SS, Fiaux K, Guimarães LC, Bakhtiar SM, Pereira U, Almeida SS, Abreu VA, Rocha FS, Dorella FA, Miyoshi A, Silva A, Azevedo V, Tauch A **Genome sequence of *Corynebacterium pseudotuberculosis* biovar equi strain 258 and prediction of antigenic targets to improve biotechnological vaccine production**. J Biotechnol. 2012 Nov 28
 9. Adriana R Carneiro, Rommel T Ramos, Hivana Dall'Agnol, Anne C Pinto, Siomar C Soares, Anderson R Santos, Luis C Guimarães, Sintia S Almeida, Rafael A Baraúna, Diego A Graças, Luciano C Franco, **Amjad Ali**, Syed S Hassan, Catarina I Nunes, Maria Silvanira Barbosa, Karina K Fiaux, Flávia F Aburjaile, Eudes G Barbosa, Syeda M Bakhtiar, Daniella Vilela, Felipe Nóbrega, Adriana L Santos, Marta Sofia P. Carepo, Vasco Azevedo, Maria Paula Cruz Schneider, Vivian Helena Pellizari, and Artur Silva. **Genome Sequence of *Exiguobacterium antarcticum* B7, Isolated from a Biofilm in Ginger Lake, King George Island, Antarctica**. J. Bacteriol. 2012. doi:10.1128/JB.01791-12
 10. Hassan SS, Schneider MP, Ramos RT, Carneiro AR, Ranieri A, Guimarães LC, **Amjad Ali**, Bakhtiar SM, Pereira Ude P, Dos Santos AR, Soares Sde C, Dorella F, Pinto AC, Ribeiro D, Barbosa MS, Almeida S, Abreu V, Aburjaile F, Fiaux K, Barbosa

- E, Diniz C, Rocha FS, Saxena R, Tiwari S, Zambare V, Ghosh P, Pacheco LG, Dowson CG, Kumar A, Barh D, Miyoshi A, Azevedo V, Silva A. **Whole-Genome Sequence of *Corynebacterium pseudotuberculosis* Strain Cp162, Isolated from Camel.** J Bacteriol. 2012 Oct;194(20):5718-9
11. Pethick. PE, Lainson. AF Raja Yaga, Flockhart A, Smith DGE, Donachie W, Louise T. Cerdeira, Silva A, Bol E, Thiago S. Lopes, Maria S. Barbosa, Pinto AC, dos Santos AR, Soares SC, Almeida SS, Guimarães LC, Aburjaile, EF Abreu VAC, Ribeiro D, **Amjad Ali**, Bakhtiar SM, Dorella FA, Carneiro AR, Ramos RTJ, Rocha ES, Schneider MPC, Miyoshi A, Azevedo V, and Fontaine MC. **Complete Genome Sequences of *Corynebacterium pseudotuberculosis* Strains 3/99-5 and 42/02-A, Isolated from Sheep in Scotland and Australia, Respectively.** J. Bacteriol. Sep. 2012 194:4736-4737; doi:10.1128/JB.00918-12
 12. Pethick. PE, Lainson. AF Raja Yaga, Flockhart A, Smith DGE, Donachie W, Louise T. Cerdeira, Silva A, Bol E, Thiago S. Lopes, Maria S. Barbosa, Pinto AC, dos Santos AR, Soares SC, Almeida SS, Guimarães LC, Aburjaile, EF Abreu VAC, Ribeiro D, Karina K. Fiaux,d Carlos A. A. Diniz, Barbosa EGV, Pereira UP, Hassan SS, **Amjad Ali**, Bakhtiar SM, Dorella FA, Carneiro AR, Ramos RTJ, Rocha ES, Schneider MPC, Miyoshi A, Azevedo V, and Fontaine MC. **Complete Genome Sequence of *Corynebacterium pseudotuberculosis* Strain 1/06-A, Isolated from a Horse in North America.** J. Bacteriol. **August 2012** vol. 194 no. 16. doi: 10.1128/ JB.00922-12
 13. Lopes T, Silva A, Thiago R, Carneiro A, Dorella FA, Rocha FS, Dos Santos AR, Lima AR, Guimarães LC, Barbosa EG, Ribeiro D, Fiaux KK, Diniz CA, de Abreu VA, de Almeida SS, Hassan SS, **Amjad Ali**, Bakhtiar SM, Aburjaile FF, Pinto AC, Soares Sde C, Pereira Ude P, Schneider MP, Miyoshi A, Edman J, Spier S, **Azevedo V.**J Bacteriol. 2012 Jul;194(13):3567-8.**Complete Genome Sequence of *Corynebacterium pseudotuberculosis* Strain Cp267, Isolated from a Llama.** J Bacteriol. 2012 Jul;194(13):3567-8.
 14. D'Afonseca V, Soares SC, **Amjad Ali**, Santos AR, Pinto AC, Magalhães AAC, Faria CJ, Barbosa E, Guimarães LC, Eslabão M, Almeida SS, Abreu VAC, Zerlotini A, Carneiro AR, Cerdeira LT, Ramos RTJ, Hirata Jr R, Mattos-Guaraldi AL, Trost E, Tauch A, Silva A, Schneider MP, Miyoshi A, Azevedo V, **Reannotation of the**

***Corynebacterium diphtheriae* NCTC13129 genome as a new approach to studying gene targets connected to virulence and pathogenicity in diphtheria,** Dove press 2012 v:4 P 1 -13.

15. R. T. J, Carneiro., D'Afonseca, V., Silva, A, **Amjad Ali**, Pinto, A. C., A.R. Santos, R, Aryanne A. M. C. Rocha, De' bora O. Lopes⁴, Dorella, F. A, Luis G. C. Pacheco, Marci' lia P. Costa, Meritxell Z. Turk, Nubia Seyffert, Pablo M. R. O. Moraes, Soares, S. C, Almeida, S. S, Thiago L. P. Castro, Abreu, V. A. C, Eva Trost, Jan Baumbach, Andreas Tauch⁶, Maria Paula C. Schneider, John McCulloch, Louise T. Cerdeira, Rommel T. J. Ramos, Zerlotini A, Dominitini A, Resende DM, Elisa^ngela M. Coser, Luciana M. Oliveira Ferro, Ortega JM, Luciano V. Paiva¹⁷, Luiz R. Goulart, Almeida JF, Maria Ine^s T. Ferro, Newton P. Carneiro, Paula R. K., Santuza M. R. Teixeira, Miyoshi, A., Guilherme C. Oliveira, Azevedo, V., **Evidence for Reductive Genome Evolution and Lateral Acquisition of Virulence Functions in Two *Corynebacterium pseudotuberculosis* Strains.** **PLoS ONE** April 2011; 6: e18551.
16. Cerdeira LT, Schneider MP, Pinto AC, de Almeida SS, Dos Santos AR, Barbosa EG, **Amjad Ali**, Aburjaile FF, de Abreu VA, Guimarães LC, Soares Sde C, Dorella FA, Rocha FS, Bol E, Gomes de Sá PH, Lopes TS, Barbosa MS, Carneiro AR, Jucá Ramos RT, Coimbra NA, Lima AR, Barh D, Jain N, Tiwari S, Raja R, Zambare V, Ghosh P, Trost E, Tauch A, Miyoshi A, **Azevedo V**, Silva A. **Complete Genome Sequence of *Corynebacterium pseudotuberculosis* Strain CIP 52.97, Isolated from a Horse in Kenya.** *J Bacteriol.* 2011 Dec;193(24):7025-6.
17. Barh D, Tiwari S, Jain N, **Amjad Ali**, Santos AR, Amarendra Narayan Misra, Vasco Azevedo, and Anil Kumar, **In Silico Subtractive Genomics for Target Identification in Human Bacterial Pathogens.** *Drug Development Research* 72 (2011).
18. Anderson Santos, **Amjad Ali**, Eudes Barbosa, Artur Silva, Anderson Miyoshi, Debmalya Barh, Vasco Azevedo. **The reverse Vaccinology, a contextual overview.** *IIOABJ, Vol. 2; Issue 4; 2011: 8_15*
19. Soares, S.C ; Rocha, Aryanne A. M. C. . Barbosa EG ,Guimarães LC. ; Almeida, S.S ; Miyoshi, A. ; Azevedo, V ; Silva, A. ; Ramos, R.T.J ; Carneiro., A. R. ; Cerdeira, L. ; **Ali, Amjad** ; SANTOS, A. ; Abreu, Vinicius A. C. ; Pinto, A.C . **Plasticidade genômica e evolução bacteriana.** *Microbiologia in Foco*, v. 4, p. 31-38, 2011

20. Cerdeira LT, Pinto AC, Schneider MP, de Almeida SS, Dos Santos AR, Barbosa EG, **Ali Amjad**, Barbosa MS, Carneiro AR, Ramos RT, de Oliveira RS, Barh D, Barve N, Zambare V, Belchior SE, Guimarães LC, de Castro Soares S, Dorella FA, Rocha FS, de Abreu VA, Tauch A, Trost E, Miyoshi A, **Azevedo V**, Silva A.. **Whole-Genome Sequence of *Corynebacterium pseudotuberculosis* PAT10 Strain Isolated from Sheep in Patagonia, Argentina**. J. Bacteriol. 2011;193 6420-6421

21. Baig SM, Koschak A, Lieb A, Gebhart M, Dafinger C, Nürnberg G, **Amjad Ali**, Ahmad I, Sinnegger-Brauns MJ, Brandt N, Enge J, Matteo E Mangoni, M. Farooq M, Khan UH, Nürnberg P, Striessnig J & Bolz HJ. **Loss of Cav1.3 (CACNA1D) function in a human channelopathy with bradycardia and congenital deafness**, *Nature Neuroscience* 2010 doi: 10.1038/nn.2694

22. Barh D, Neha Jain, Sandeep Tiwari, Vivian D'Afonseca, Liwei Li, **Amjad Ali**, Anderson Rodrigues Santos, Luís Carlos Guimarães, Siomar de Castro Soares³, Anderson Miyoshi, Atanu Bhattacharjee⁵, Amarendra Narayan Misra, Artur Silva, Anil Kumar, Azevedo V. **A novel comparative genomics analysis for common drug and vaccine targets in *Corynebacterium pseudotuberculosis* and other CMN group of human pathogen**. *Chem Biol Drug Des* 2011.

23. Silva, A. Schneider, M. P. C. Cerdeira, L. Barbosa, M. S. Ramos, R. T. J, Carneiro, A.R. Santos, R. Lima, M. D'Afonseca, V. Almeida, S. S., Santos, A. R.; Soares, S. C. ; Pinto, A. C., **Amjad, Ali.**, Dorella, F. A. ; Rocha, F. ; Abreu, V. A. C. ; Shpigel, N. Miyoshi, A., Azevedo, v., **Complete Genome Sequence of *Corynebacterium pseudotuberculosis* I-19, strain isolated from Israel Bovine mastitis**. Journal of Bacteriology (Print) **JCR**, p. 1-1, 2010.

24. Farooq, M., Troelsen, J.T., Boyd, M., Hansen, L., Eiberg, H., Hussain, M.S., Rehman, U.S., Azhar, A., **Amjad Ali.**, Bakhtiar, SM., Tommerup, N., Baig, S.M. and Klaus, W.K. **Preaxial polydactyly/triphalangeal thumb is associated with changed transcription factor binding affinity in a family with a novel point mutation in the long range cis-regulatory element ZRS**. Eur J Hum Genet. Epub 2010 Jan 13.

25. Rasool, M., Schuster, J., Aslam, M., Tariq, M., Ahmad, I., **Amjad Ali.**, Entesarian, M., Dahl, N., Baig, S.M. **A novel missense mutation in the EDA gene associated with**

X-linked recessive isolated hypodontia. J Hum Genet, 2008; 53(10): 894-8. Epub 2008 Aug 9.

Book Chapters

1. Vasco Azevedo, Vinicius Abreu, Sintia Almeida, Anderson Santos, Siomar Soares, **Amjad Ali**, Anne Pinto, Aryane Magalhães, Eudes Barbosa, Rommel Ramos, Louise Cerdeira, Adriana Carneiro, Paula Schneider, Artur Silva and Anderson Miyoshi (2011). **Whole Genome Annotation: In Silico Analysis**, Bioinformatics - Trends and Methodologies, Mahmood A. Mahdavi (Ed.), ISBN: 978-953-307-282-1, InTech, Available from: <http://www.intechopen.com/articles/show/title/whole-genome-annotation-in-silico-analysis>

Conferences/Proceedings

- **Amjad Ali**, D'afonseca V, Santos ARD, Pinto AC, Magalhães AAC, Faria CJ, Barbosa E, Dorella FA, Pacheco LGC, Almeida SS, Soares SC, Abreu VAC, Silva A, Moore R, Miyoshi A, Azevedo V. **Identification, Characterization and In silico comparisons of *Corynebacterium pseudotuberculosis* Pathogenicity Islands (PAIs) using PIPS.** 6th International conference of Brazilian association for bioinformatics and computational biology, X-meeting 15th-18th Nov. 2010. Ouro Preto MG Brazil.
- D'afonseca V, **Ali A**, Santos ARD, Pinto AC, Magalhães AAC, Faria CJ, Barbosa E, Dorella FA, Pacheco LGC, Almeida SS, Soares SC, Abreu VAC, Silva A, Moore R, Miyoshi A, Azevedo V. **Computational approach to study the genome reduction and life style of *Corynebacterium pseudotuberculosis* strains.** 6th International conference of Brazilian association for bioinformatics and computational biology, X-meeting 15th-18th Nov. 2010. Ouro Preto MG Brazil.

Training/ Courses

- Platform for Next Generation Sequencing "NGS-SOLiD" Sequencing, Assembly and Annotation of Bacterial Genomes (90hrs). Aug 30 to Sep 10, 2010, UFPA, Belem, Brazil.
- From bench to bedside: bacterial pathogenesis and its impact on the clinical outcome of infections. School of veterinary sciences, UFMG, Brazil 2011.
- Comparative Genomic Course and Training, Center for Biological Sequence Analysis Technical University of Denmark (Sep. 2011).
- Research Ethic certificate course (96hrs) Aga Khan University /John Hopkins USA. (Aug. 2009)
- Molecular Detection and Genotyping of HCV/ HBV by PCR methods and Quality Assurance Issues. National Institute for Biotechnology and Genetic Engineering NIBGE, Faisalabad.
- Nuclear and other Advance Techniques in Agricultural and Biological Sciences at National Institute for Agriculture and Biology, Faisalabad (2006)
- Bio-Rad Laboratories Inc. USA (life science research) Chemical-House Lahore (Mar-Dec. 2006)

Languages Skills

English, Urdu and Portuguese

References

Prof. Dr. Vasco Azevedo
Depto de Biologia Geral, ICB/UFMG.
Av. Antonio Carlos, 6627. Pampulha,
Belo Horizonte, Minas Gerais, Brazil.
CP 486 CEP 31270-901
Tel/FAX 005531 3409 2610
Email: vascoariston@gmail.com

Prof. Dr. Anderson Miyoshi
Depto de Biologia Geral, ICB/UFMG,
Av. Antonio Carlos, 6627. Pampulha,
Belo Horizonte, Minas Gerais, Brazil.
CP 486 CEP 31270-901
Tel/FAX 005531 3409 2610
Email: miyoshi@icb.ufmg.br

IX. APPENDIX

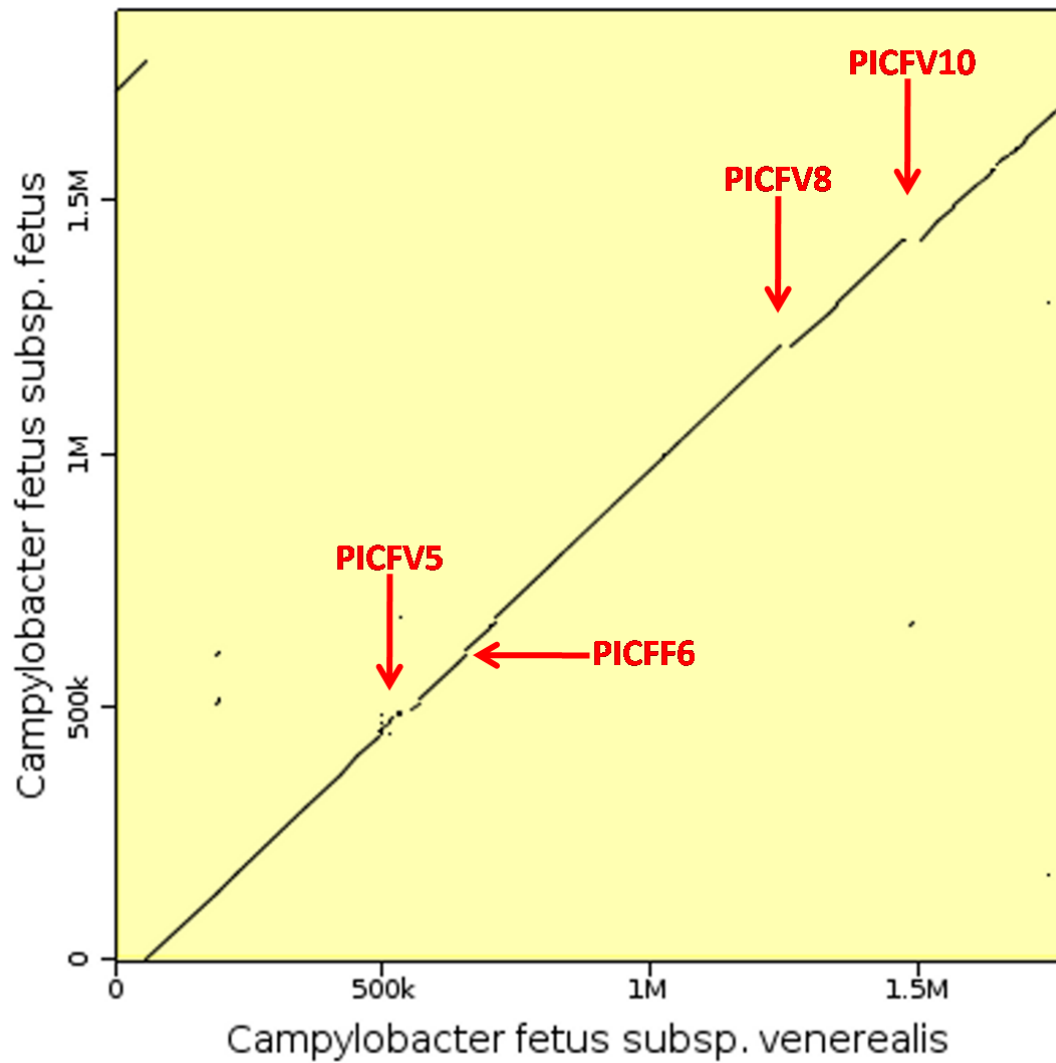
IX.I Chapter 1. *Corynebacterium*

Virulence factors	Related genes	<i>C. diphtheriae</i> NCTC 13129 NC 002935	<i>C. efficiens</i> YS-314 NC 004369	<i>C. glutamicum</i> <i>R</i> NC 009342	<i>C. jeikeium</i> K411 NC 007164	<i>C. pseudotuberculosis</i> FRC41 NC_014329
Adherence						
Collagen-binding protein	<i>cbpA</i>				jk0461	
SpaA-type pili	<i>spaA</i>	DIP2013	CE2737, CE2455, CE2457			
	<i>srtA</i>	DIP2012	CE2738			cpfrc_01905
	<i>spaB</i>	DIP2011				
	<i>spaC</i>	DIP2010				cpfrc_01901, cpfrc_01902, cpfrc_01903
SpaD-type pili	<i>srtB</i>	DIP0233				cpfrc_01875
	<i>spaD</i>	DIP0235		cgR_2789	jk1701	cpfrc_01874
	<i>srtC</i>	DIP0236			jk1700	cpfrc_01873
	<i>spaE</i>	DIP0237			jk1699	
	<i>spaF</i>	DIP0238				
SpaH-type pili	<i>spaI</i>	DIP2223				cpfrc_01904
	<i>srtE</i>	DIP2224		cgR_2790, cgR_2793		
	<i>srtD</i>	DIP2225	CE2454, CE2456			
	<i>spaH</i>	DIP2226		cgR_2791		
	<i>spaG</i>	DIP2227				
Surface-anchored pilus proteins	<i>sapA</i>	DIP2066			jk1702	cpfrc_01870
	<i>sapD</i>	DIP0443			jk1856	
	<i>sapE</i>				jk0007	
Iron uptake						
ABC transporter	<i>fagC</i>	DIP1059	CE0684	cgR_0600	jk1774	cpfrc_00030
	<i>fagB</i>	DIP1060	CE0685	cgR_0601	jk1773	cpfrc_00031
	<i>fagA</i>	DIP1061	CE0686	cgR_0602	jk1772	cpfrc_00032
	<i>fagD</i>	DIP1062	CE0687	cgR_2963	jk1776	cpfrc_00033
ABC-type heme transporter	<i>hmuT</i>	DIP0626		cgR_0462	jk0316	cpfrc_00455
	<i>hmuU</i>	DIP0627	CE0694	cgR_0463	jk0317	cpfrc_00456
	<i>hmuV</i>	DIP0628	CE0693	cgR_0464	jk0318	cpfrc_00457
ciu iron uptake and siderophore biosynthesis system	<i>ciuA</i>	DIP0582			jk1296	cpfrc_00987
	<i>ciuB</i>	DIP0583			jk1295	cpfrc_00988
	<i>ciuC</i>	DIP0584			jk1294	cpfrc_00989
	<i>ciuD</i>	DIP0585			jk1293	cpfrc_00990
	<i>ciuE</i>	DIP0586				cpfrc_00991
Siderophore-depenedent iron uptake system	<i>irp6A</i>	DIP0108			jk0561	
	<i>irp6B</i>	DIP0109			jk0560	
	<i>irp6C</i>	DIP0110			jk0559	
Regulation						
Diphtheria toxin repressor DtxR	<i>dtxR</i>	DIP1414	CE1812	cgR_1750	jk1097	cpfrc_01219
Toxin						
Diphtheria toxin (DT)	<i>tox</i>	DIP0222				
Phospholipase D	<i>pld</i>					cpfrc_00029

Additional TABLE 1. *Corynebacterium* species comparative pathogenomics.

Modified from <http://www.mgc.ac.cn/cgi-bin/VFs/>

IX.II Chapter 2. *Campylobacter*



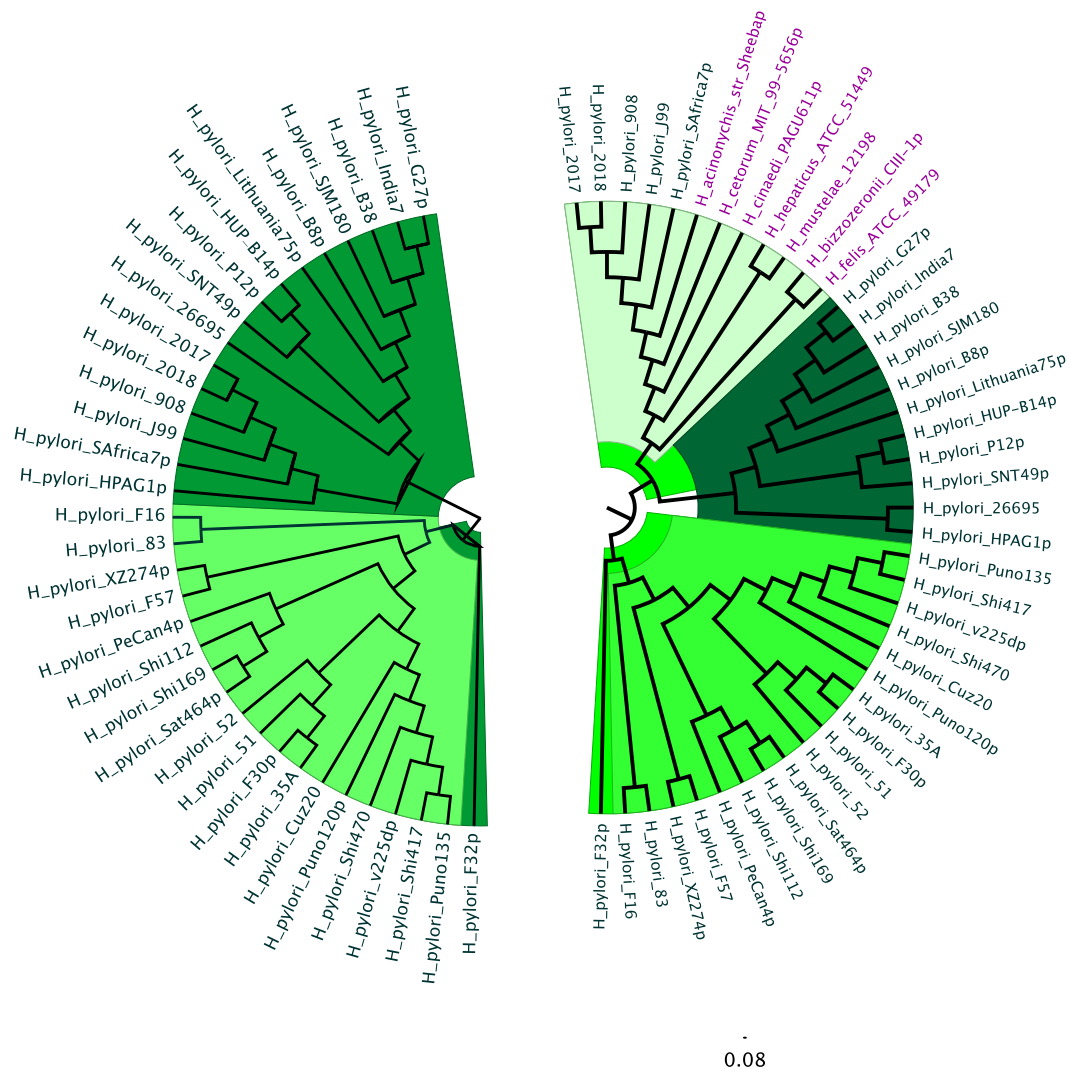
ADDITIONAL FIGURE 1. Synteny plot showing pathogenicity island insertion regions on the genome sequences of *C. fetus* subspecies.

Organism	Cytoplasm	Membrane	Secreted	PSE	Total
<i>Cff</i>	1174 (68%)	244 (14%)	123 (7%)	178 (10%)	1719
<i>Cfv</i>	1472 (70%)	274 (13%)	142 (7%)	204 (10%)	2092
<i>C. concisus</i>	1380 (66%)	280 (13%)	181 (9%)	239 (11%)	2080
<i>C. curvus</i>	1271 (66%)	154 (13%)	171 (9%)	235 (12%)	1931
<i>C. hominis</i>	1148 (68%)	244 (14%)	104 (6%)	191 (11%)	1687
<i>C. lari</i>	1067 (69%)	180 (12%)	103 (6%)	196 (12%)	1546
<i>C. J1221</i>	1291 (70%)	231 (12%)	121 (6%)	195 (11%)	1838

FIGURE 5 DATA IN TABLE: Subcellular localization of *Campylobacter*

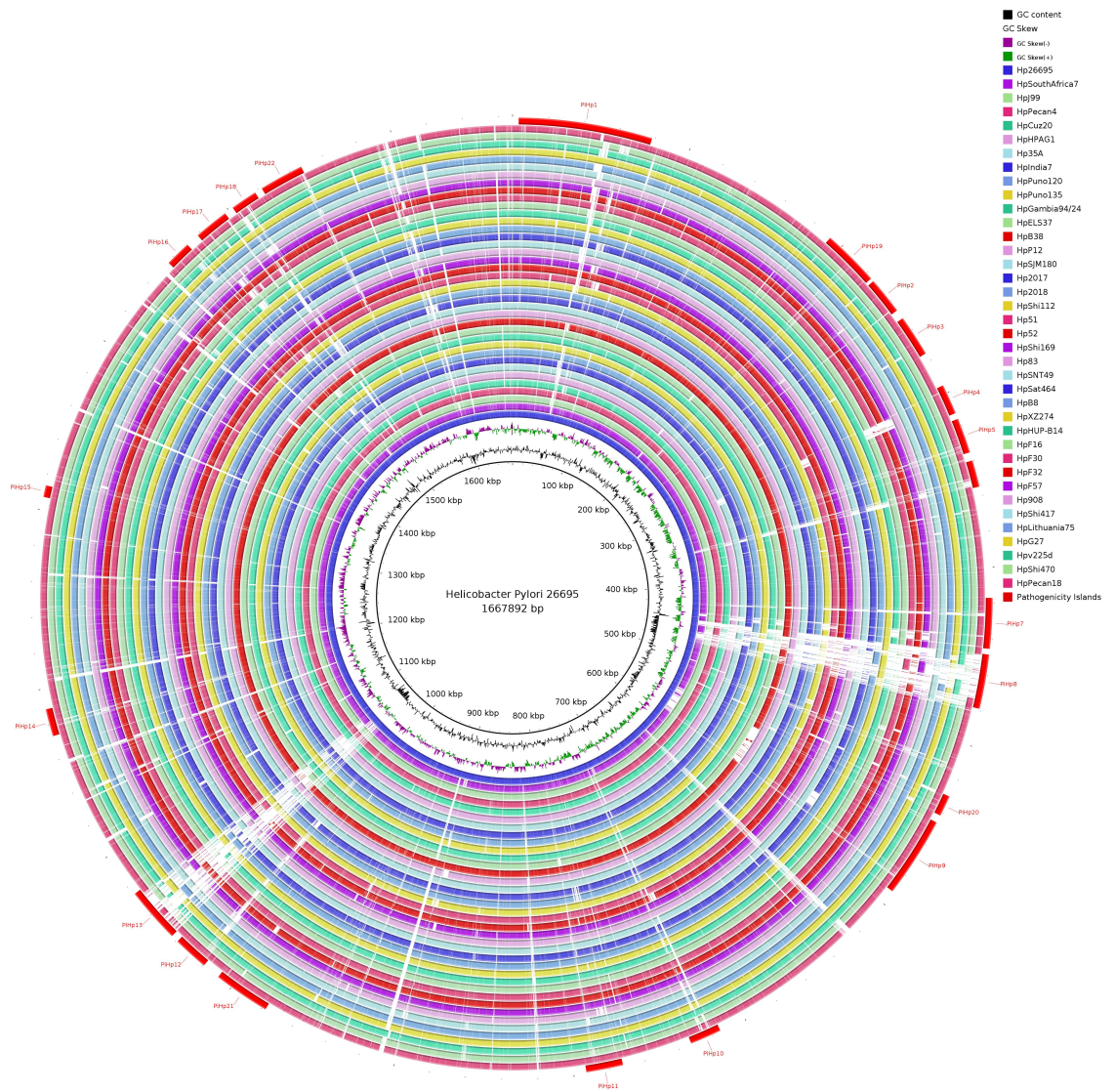
species proteins: *Cfv*, *C. fetus* subsp. *venerealis* NCTC 10354T; *Cff*, *C. fetus* subsp. *fetus* 82-40; *CJ1221*, *C. jejuni* RM1221; PSE, potential surface exposed. The numbers (%) indicates the proteins in corresponding locations in the cells.

IX.III Chapter 3. *Helicobacter*



ADDITIONAL FIGURE 1. Phylogenetic tree based on 16S rRNA gene divergence.

Both the genus (right) *Helicobacter* and *H. pylori* strains (left) are analysed.



ADDITIONAL FIGURE 3. *H. pylori* Genome comparisons, genome plasticity analysis and pathogenicity islands predicted. *H. pylori* strain 26695 is used as a reference. The genome plasticity analysis in *H. pylori* genomes and putative potential pathogenicity islands are predicted (Program used: PIPS and BRIG).