Tesis de Doctorado en Biología (PEDECIBA)

Subárea Biología Celular y Molecular

Desarrollo y aplicación de herramientas computacionales para el análisis taxonómico y patogenómico de procariotas

October 11, 2016

Unidad de Bioinformática
Institut Pasteur Montevideo

Autor

Gregorio iraola

Orientador

Hugo naya

Tribunal

Elena Fabiano (Presidente)
Alejandro Buschiazzo (Vocal)
Héctor Romero (Vocal)

Prefacio

En esta Tesis se presentan de forma unificada las actividades de investigación llevadas a cabo en la Unidad de Bioinformática (en colaboración con otras instituciones locales y extranjeras) entre los años 2010 y 2016. Estas actividades comenzaron en el marco de mi Maestría en Bioinformática (PEDECIBA), donde desarrollé algunas aproximaciones para la predicción de patogenicidad a partir del análisis de genomas bacterianos. Estos resultados dieron lugar a una línea de investigación dedicada al estudio genómico de microorganismos y a optar por el pasaje al programa de Doctorado en Biología, subárea Biología Celular y Molecular (PEDECIBA) en 2014. En estos seis años de formación de posgrado he continuado con el desarrollo y aplicación herramientas de biología computacional para el análisis de genomas procariotas, con el objetivo específico de responder preguntas generales y particulares acerca de la clasificación taxonómica de los procariotas y su patogenicidad utilizando datos producidos por tecnologías de secuenciación masiva.

El hilo conductor de la Tesis es el estudio de genomas procariotas por medios computacionales aunque, debido a la diversidad de temas específicos abordados, la misma ha sido dividida en tres partes y un total de diez capítulos. Cada capítulo representa en su totalidad un artículo científico ya publicado o en vías de publicación en revistas científicas internacionales. La **parte uno** consta de tres capítulos que describen herramientas de análisis desarrolladas para la predicción de patogenicidad y la clasificación taxonómica de procariotas a partir de sus genomas. La **parte dos** consta de cuatro capítulos centrados en diferentes aspectos de la biología del género *Campylobacter*, desde la descripción taxonómica de nuevas especies al estudio de la epidemiología y patogenicidad de especies ya conocidas. Finalmente, en la **parte tres** confluyen trabajos puntuales que se han desarrollado en la misma línea de investigación, particularmente en el estudio transcripcional de leptospiras formadoras de biofilms y la producción y análisis de genomas de casos clínicos de *Mycobacterium tuberculosis* en Uruguay.

Contenidos

Lista de Figuras

Lista de Figuras Suplementarias

Lista de Tablas

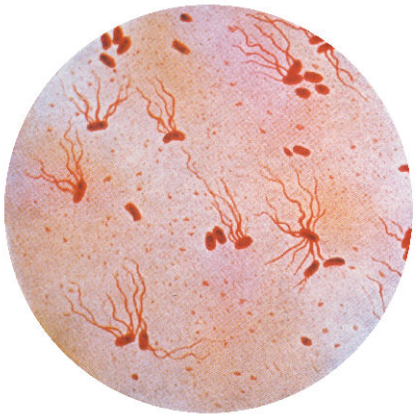
Lista de Tablas Suplementarias

Parte I

Predicción de fenotipos a partir de genomas bacterianos

(Predicting phenotypes from bacterial genomes)

1 Genome-based prediction of bacterial pathogenicity in humans



Citation:

Iraola G, Vázquez G, Spangenberg L, Naya H* (2012) **Reduced set of virulence-related genes allows high accuracy prediction of bacterial pathogenicity in humans** *PLoS ONE*. 7(8): e42144.

* Corresponding author

1.1 Abstract

Although there have been great advances in understanding bacterial pathogenesis, there is still a lack of integrative information about what makes a bacterium a human pathogen. The advent of high-throughput sequencing technologies has dramatically increased the amount of completed bacterial genomes, for both known human pathogenic and nonpathogenic strains; this information is now available to investigate genetic features that determine pathogenic phenotypes in bacteria. In this work we determined presence/absence patterns of 814 different virulence-related genes among more than 600 finished bacterial genomes from both human pathogenic and non-pathogenic strains, belonging to different taxonomic groups (i.e: Actinobacteria, Gammaproteobacteria, Firmicutes, etc.). An accuracy of 95% using a cross-fold validation scheme with in-fold feature selection is obtained when classifying human pathogens and non-pathogens. A reduced subset of highly informative genes (120) is presented and applied to an external validation set. The statistical model was implemented in the BacFier v1.0 software, that displays not only the prediction (pathogen/non-pathogen) and an associated probability for pathogenicity, but also the presence/absence vector for the analyzed genes, so it is possible to decipher the subset of virulence genes responsible for the classification on the analyzed genome. Furthermore, we discuss the biological relevance for bacterial pathogenesis of the core set of genes, corresponding to eight functional categories, all with evident and documented association with the phenotypes of interest. Also, we analyze which functional categories of virulence genes were more distinctive for pathogenicity in each taxonomic group, which seems to be a completely new kind of information and could lead toi mportant evolutionary conclusions.

1.2 Introduction

Several factors, including globalization and sanitation conditions, have been shaping the world's landscape of infectious diseases over the years. In developed countries, 90 percent of documented infections

in hospitalized patients are caused by bacteria. These cases probably show only a small proportion of the actual number of bacterial infections occurring in the entire population, and they usually represent the most severe cases. In developing countries, a variety of bacterial infections often provoke a devastating effect on the inhabitantsÕ health. The World Health Organization (WHO) has estimated that each year, 1,3 million people die of tuberculosis, 0,2 million die of pertussis and 0,1 million die of syphilis. Diarrheal diseases, many of which are of bacterial etiology, are the second leading cause of death in the world (after cardiovascular diseases), killing 2,5 million people annually. This scenario evidences that even today, infectious diseases are a permanent threat for human health around the world.

Understanding the biology of the causative agents of these diseases has been a permanent challenge since the beginning of bacteriology. Nowadays, the mechanisms involved in the virulence (defined as the relative capacity of a microbe to cause damage in a host) of pathogenic bacteria are widely studied in clinical bacteriology, but the advent of new technologies has enabled their study from different perspectives. In this context, bacterial genomics have greatly contributed to the better understanding of pathogenicity due to the possibility of generating and comparing whole genome sequences. The onset of this discipline started with the automation of Sanger sequencing chemistry and the completion of *Haemophilus influenzae* and *Mycoplasma genitallium* genomes [1, 2] in the mid-1990s; since then, projects to sequence the genomes of a large number of organisms were undertaken by means of this method [3–5]. However, during the last decade, to cover the increasing sequencing demands, new non-Sanger high-throughput sequencing systems have been developed under the name of "second generation" or "next-generation" sequencing technologies [6, 7]. These developments have significantly reduced the cost and simultaneously increased the speed of DNA sequencing. In this sense, the great majority of organisms whose genomes have been sequenced so far are bacteria, with 1505 complete and published genome sequences and 6037 ongoing projects.

Comparative genomics, including comparison at the DNA, transcriptome, and proteome levels, have emerged as a key to give a biological sense to all this massive information. Focused on improving

the knowledge on pathogenicity determinants two bioinformatic approaches have been used, based on two complementary explanations for bacterial pathogenesis. On the one hand, pathogenicity has been related to amino acid substitutions which lead to modified protein structures, and probably modified functions [8–10]. In this case, a particular gene shared by a human pathogenic species and a non-pathogenic species, could be causing a pathogenic phenotype in the first one, determined by non-synonymous mutations that modify key amino acids and alter protein function. Based on this, our group has recently published a method that detects variable regions inside protein sequences which can be potentially related to pathogenicity [11].

On the other hand, trying to give an integrative view of bacterial pathogenicity prediction from a bioinformatics perspective, in this work we exploit an alternative explanation for bacterial pathogenicity. Pathogenicity has been attributed to the presence or absence of genes which confer particular pathogenic phenotypes, like toxins [12]. In this case, these genes would be present in pathogenic species but absent in non-pathogenic ones. The most widely spread approach to evaluate this is the pairwise comparison between genomes of pathogenic and non-pathogenic bacteria or even multiple comparisons between different strains of the same species [13–15]. These kinds of approaches can give information regarding the presence or absence of genes involved in pathogenicity of a particular species or even a genus. However, it is difficult to extrapolate this information to higher taxonomic levels, which keeps us from drawing conclusions about general features that are determining bacterial pathogenicity.

For this reason, our motivation was: i) try to identify presence/absence patterns of virulence-related genes which could explain the pathogenic phenotype of bacteria at higher taxonomic levels than species or genus, ii) discuss the biological significance of those genes giving an integrative view of genetic determinants of bacterial pathogenicity, iii) use this information to develop a machine learning model to classify bacterial genomes into human pathogens and non-pathogens and iv) implement this model in a software that can be used to predict pathogenicity in the upcoming sequenced bacterial genomes. The last two points are particularly interesting because a statistical model implemented in an easy-to-use software, capable of predicting

bacterial pathogenicity based on genomic information, can be helpful for practical purposes. For example, in food or pharmaceutical industries it is essential to know the pathogenic potential of bacterial strains used in bioengineering.

1.3 Results and Discussion

The idea that bacterial species can be effectively grouped into human pathogens and non-pathogens based on their virulence genes composition, arises from preliminary results that indicated differential patterns in presence or absence of these kind of genes among both groups (human pathogens and non-pathogens).

All finished and annotated genomes of human pathogenic and non-pathogenic bacteria were used to perform a presence/absence analysis over 814 groups of orthologous genes belonging to 8 functional categories (toxins, two-component systems, ABC transporters, motility, flagellar assembly, LPS biosynthesis, secretion systems and chemotaxis), in order to determine which ones are strongly related to pathogenicity in different bacterial taxonomic groups (Actinobacteria, Alphaproteobacteria, Betaproteobacteria, Bacteroidetes/Chlorobi, Chlamydiae/Verrucomicrobia, Deltaproteobacteria, Epsilonproteobacteria, Firmicutes, Gammaproteobacteria, Spirochaetes, etc.). Fig. 1.1 shows phylogenetic relations and the proportion of pathogenic and non-pathogenic organisms in studied taxa.

The analysis was accomplished by calculating the frequency of genes belonging to each functional category in pathogenic and non-pathogenic species of each taxon. The assumed null hypothesis was that, if a certain gene is not related to pathogenicity, its frequency would not be biased towards pathogenic or non-pathogenic species; furthermore, it would be almost equally distributed within both classes. Genes presenting a high frequency among pathogens and a low frequency in non-pathogens are probably contributing to a pathogen-related phenotype, for example genes coding for toxins. Conversely, a gene that presents low frequency among pathogens and high frequency in non-pathogens could be indicating the loss of genes coding for redundant functions. For example, proteins that transport certain molecules across membranes, which are essential for a free-

living style, are often dispensable when bacteria are well-adapted to the environment inside their hosts. The frequency distribution of ABC transporter genes in Alphaproteobacteria and Gammaproteobacteria clearly exemplifies this situation. Fig. 1.2 shows the frequency of each gene in pathogenic and non-pathogenic organisms. Points falling on the diagonal line represent genes whose frequency is balanced between pathogens and nonpathogens. Points closer to the Y axis are more represented in non-pathogens and points closer to the X axis are more frequent in pathogens. As it is shown in this figure, ABC genes are strongly related to non-pathogenic species in Alphaproteobacteria, while there are overrepresented in pathogenic species in Gammaproteobacteria.

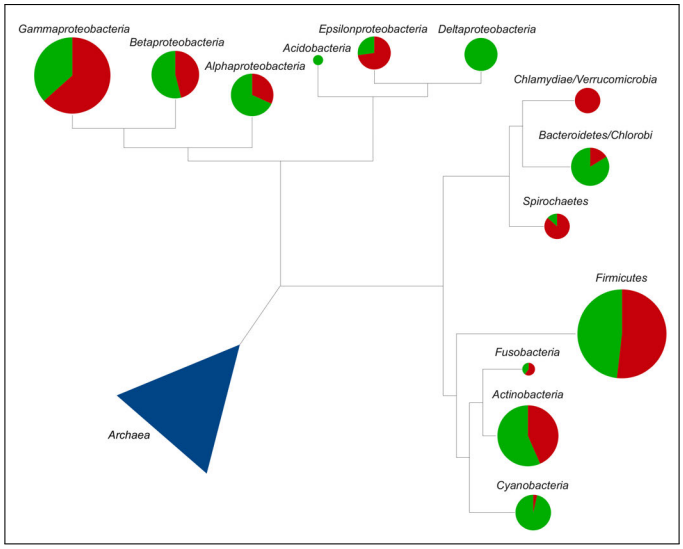


Figure 1.1: Phylogenetic relations of bacterial groups used in this work. Chart sizes are proportional to the number of genomes present in each taxonomic group. The percentage of pathogenic organisms is shown in red and green is used for non-pathogenic.

As shown in Fig. 1.3 the number of present genes is highly variable among classes (pathogens and non-pathogens) and even between taxonomic groups. Moreover, a great number of these present genes, belonging to the 8 functional categories, presented a frequency bias towards either pathogenic or non-pathogenic species (Fig. 1.4), deviating from the proposed null hypothesis. These findings supported the

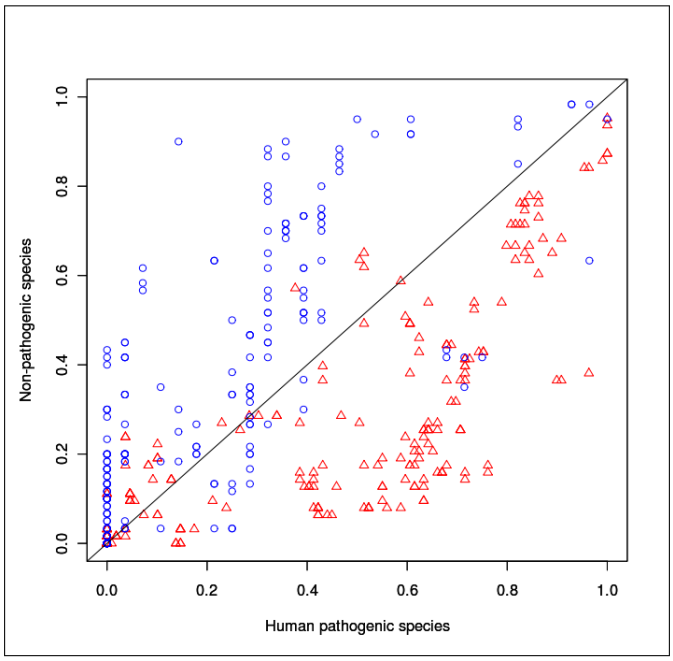


Figure 1.2: Frequency distribution of ABC transporter genes in Alphaproteobacteria and Gammaproteobacteria. For each gene, abcisse value is the number of pathogenic strains inside a certain taxonomic group in which it is present, divided by the total number of pathogenic strains inside the taxonomic group. The ordinate value is the same but for the non-pathogenic strains inside the group. Blue circles show that genes coding for ABC transporters are more frequent in pathogenic species of Gammaproteobacteria than in non-pathogenic species of this group. The opposite pattern is observed for Alphaproteobacteria in red triangles.

idea that presence/absence patterns of virulence-related genes are informative enough to discriminate between human pathogenic and non-pathogenic bacterial species (Tab. 1.1), indicating that this data can be used to construct a classification model based on highly significant biological information.

1.3.1 Classification model. We used a machine learning approach based on a cross-validation scheme with in-fold feature selection together with a linear Support Vector Machine (SVM) classifier. Preliminary models were constructed using the whole 814 set of genes, but the number of genes was systematically reduced by means of a feature selection process. The definitive model included the first 120 genes

ranked by their significance for classification (Tab. 1.6). However, since the number of variables is still high, problems associated with chance correlation might arise. For these reason a y-randomization test was implemented. Supp. Fig. 1.1 shows the performance obtained in the test (50% accuracy), indicating the absence of chance correlation. Section "Model construction" further explains these methodologies.

Table 1.1: Statistical overview of data distribution among phyla. Statistical depression is measured as the interquartile range (IQR) in human pathogens (HP) and non-pathogens (NP).

		Class NP					Class HP				
Taxon	Purpose[1]	n	median	IQR	min	max	n	median	IQR	min	max
Actinobacteria	M	43	17.0	7.00	9	28	33	12.0	4.00	5	20
Alphaproteobacteria	M	60	28.5	13.50	0	49	28	10.0	23.00	5	37
Bacteroidetes/Chlorobi	M	26	10.5	3.75	0	15	5	8.0	4.00	7	11
Betaproteobacteria	M	28	29.5	11.50	11	47	23	39.0	25.00	14	49
Epsilonproteobacteria	M	6	17.5	9.75	7	20	16	14.0	0.25	13	20
Firmicutes	M	93	20.0	10.00	0	30	100	16.0	10.00	3	30
Gammaproteobacteria	M	63	25.0	15.00	1	47	109	43.0	24.00	9	51
Spirochaetes	M	2	20.0	6.00	14	26	13	9.0	1.00	8	14
Chlamydiae/Verrucomicrobia	T	-	-	-	-	-	14	11.0	0.00	10	12
Deltaproteobacteria	T	28	22.0	5.25	6	31	-	-	-	-	-

[1]M: used in model construction, T: used in model testing

The number of correctly/incorrectly classified genomes in the complete set was 618/30, obtaining an accuracy of 95.4%. Tab. 1.2 describes the classification performance related to all bacteria taxonomy considered in the dataset. The last column of the table indicates the classification success rate for each group considered in the taxonomy; all values were obtained using the 10- fold cross validated SVM model, not by retraining the model using only organisms of the particular taxon. The performance is preserved across the whole taxonomy, ranging from 91% in Epsilonproteobacteria, up to 100% in Bacteroidetes/Chlorobi. Mid-size groups like Betaproteobacteria, Actinobacteria and Alphaproteobacteria showed a prediction success rate similar or better than the general performance rate. Finally the Firmicutes, the biggest group, showed an excellent classification level of 97.4%. Classification performance according to class labels is shown in Tab. 1.3, the general error rate is almost equal for false positives and negatives and

the general success rate is also equal for pathogens and non-pathogens.

1.3.2 Model testing and comparison. To further test the SVM model we evaluated its performance by analyzing genomes originally not included in the dataset used to construct the model. On the one hand, we defined a Group I of 124 genomes with known labels for human pathogen or non-pathogen, originally excluded from the dataset due to reduced number of genomes per group or misrepresentation of one of the two classes. On the other hand, we defined a Group II of 232 "blind" genomes without previous information for pathogenicity.

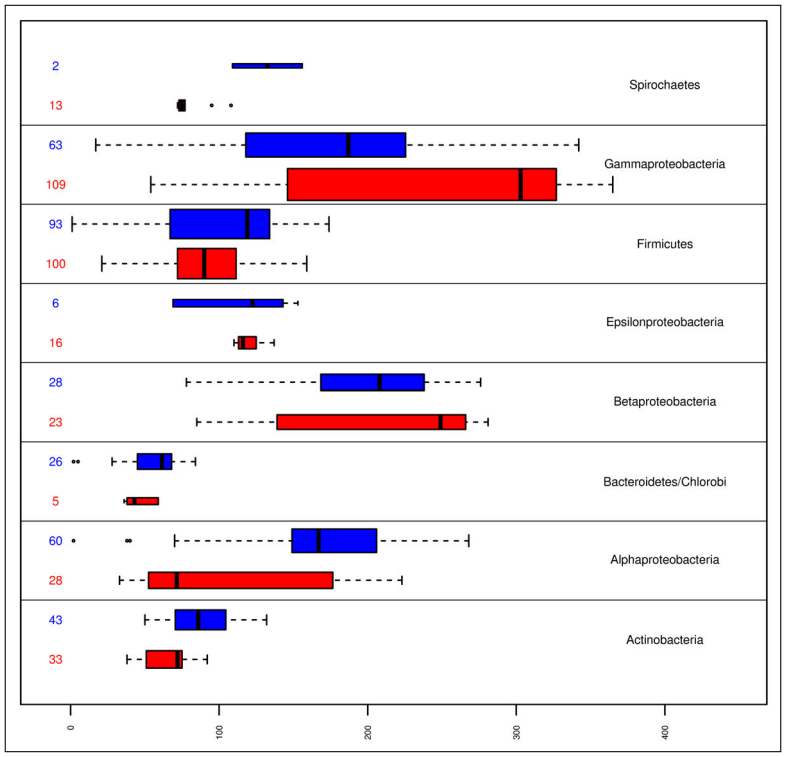


Figure 1.3: Boxplot representing the presence of genes per taxonomic group. The length of each box represent the number of genes present in both pathogenic (orange) and non-pathogenic (green) organisms for each taxonomic group considered. The number of organisms inside each group are shown leftside, this number is proportional to box width. Dark vertical lines show the median for the amount of present genes per group, box limits represent quartiles and whiskers extend to the most extreme data point which is no more than 1:5 times the interquartile range.

Group I genomes were classified with an accuracy of 98% (Tab. 1.4), even better than the average 95.4% obtained during cross-validation procedure using the original dataset. Only in two taxonomic groups (Chlamydiae/Verrucomicrobia and Fusobacteria) the model showed an accuracy lower than 100%, and in each case only one genome was misclassified. Group II genomes were previously subjected to an exhaustive bibliographic search in order to assign them to human pathogens or non-pathogens. Application of SVM model over this group resulted in 92% of average accuracy (Tab. 1.4), ranging from 87% in Epsilonproteobacteria to 100% in Deltaproteobacteria, Bacteroidetes, etc. The fact that accuracy is preserved in both test groups reaffirms the results obtained when performing the cross-validation scheme, indicating that our model is robust and the high performance in classification and prediction of human pathogens and non-pathogens is independent of the dataset used to build the model.

Table 1.2: Classification performance for each phylum used to construct the model. Inside each class are shown the number of correct and incorrect classified genomes.

		Class NP		Class HP		
	Number	Pred. NP	Pred. HP	Pred. NP	Pred. HP	Precision
Actinobacteria	76	42	1	31	2	96.05%
Alphaproteobacteria	88	60	0	26	2	97.73%
Bacteroidetes/Chlorobi	31	26	0	5	0	100%
Betaproteobacteria	51	26	2	22	1	94.12%
Epsilonproteobacteria	22	6	0	16	0	100%
Firmicutes	193	90	3	96	4	96.37%
Gammaproteobacteria	172	54	9	104	5	91.86%
Spirochaetes	15	2	0	12	1	93.33%

The SVM model was also compared to a method developed by Andreatta *et al.* [16], which is the unique tool reported so far with the same purpose of predicting bacterial pathogenicity. Andreatta *et al.* proposed a classifier for the prediction of pathogenicity restricted only to Gammaproteobacteria, considering a dataset of 155 organisms and obtaining an accuracy of 87%. This is lower than the 96.5% achieved for the same taxonomic group (using 172 organisms) with our SVM model, and even worse than the general performance of our classifier (95.4%). Furthermore, in the particular case of Gammaproteobacte-

ria, our method presented a lower error rate in misclassifying human pathogens as non-pathogens (only 1/50), than the other way around (1/15) non-pathogens classified as pathogens). This is of crucial importance in practical applications (such as for clinical or industrial purposes), since the social costs of misclassifying a pathogenic strain as non-pathogenic are usually higher than the opposite scenario.

Table 1.3: Confusion matrix showing the average classification precision across phyla.

Classified as	Pathogenic	Non-pathogenic
Pathogenic	337 (95.74%)	15 (4.59%)
Non-pathogenic	15 (4.26%)	284 (95.41%)

1.3.3 Biological interpretation. The eight pathogenicity-related functional categories investigated in this work were represented in the set of 120 genes selected for the classifier. Forty genes belonged to ABC transporters, 41 corresponded to two-component systems and chemotaxis proteins, 11 corresponded to toxins, 6 belonged to the LPS biosynthesis pathway and 22 coded for flagellar assembly proteins, motility proteins and proteins from secretion systems. We selected from each group the most distinctive genes and discussed their biological meaning considering their implications in bacterial pathogenesis (Tab. 1.5).

ABC transporters. ABC transporters are specialized proteins that function as either importers, which bring nutrients and other molecules into cells, or as exporters, which pump toxins, drugs and lipids across membranes [17]. Based on the kind of substrate ABC transporters are specific for: i) metallic cations, iron-siderophore and vitamin B12, ii) phosphate and amino acids, iii) oligosaccharides and polyol, iv) monosaccharides, v) mineral and organic ions, vi) peptides and nickel and vii) others (ABC-2). Our classification model selected those ABC transporters related to transport of metallic cations, vitamin B12, phosphate and amino acids as the most important.

It is widely known that metallic ions, are essential for prokaryotic cell physiology. The amount of these ions is not constant inside the hosts of pathogenic bacteria, and their concentration is sometimes considerably lower than needed [18]. The presence of systems implied in

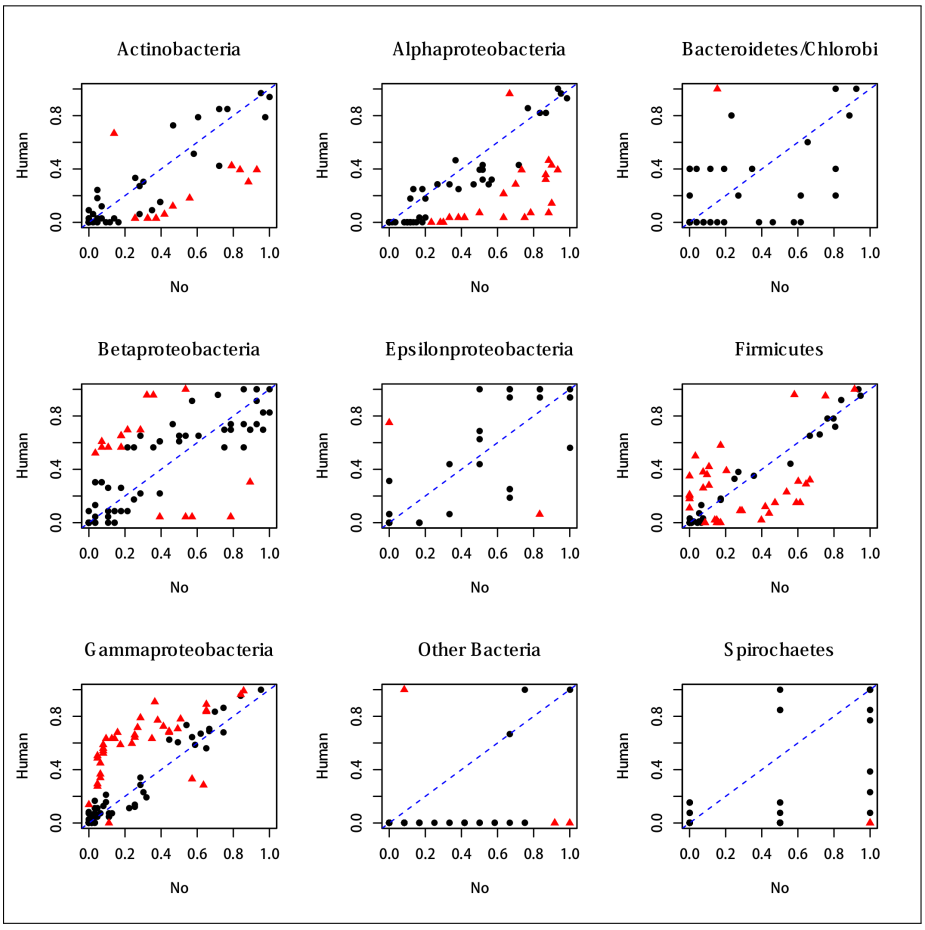


Figure 1.4: Frequencies of each of 814 genes per bacterial taxonomic group. Frequency calculation was performed for each gene as in Fig. 1.2. Red triangles show significative genes that apart from the null distribution (same frequency in pathogens and non-pathogens) by exact Fisher test, black circles are non significative genes.

metallic cations scavenging is mandatory for bacterial survival inside host cells, and it is a key feature for downstream processes like the development of pathogenic phenotypes [19].

The emergence of most pathogenic species is associated with an evolutionary transition from a free-living to a host dependent lifestyle, to a certain extent. Bacterial genomes, and especially those from pathogens, abide by the maxim "use it or leave it", where genes or even whole gene pathways are lost if their products are not essential for cell

maintenance, or can be taken from the environment [20]. Two examples are amino acid and vitamin biosynthesis pathways, which have been lost in most pathogens. In this sense, the high representation of these types of ABC systems support the idea that it is more convenient for pathogens to incorporate these compounds from the host environment than to produce them *de novo*.

Two-component systems and chemotaxis. Two-component systems (TCS) are widespread signal transduction pathways among bacteria, which play a crucial role in adaptation to fluctuating surroundings by sensing changes in environmental conditions, like those experimented during process of entry, colonization and spread [21]. Genes belonging to 9 TCS families were selected by the classifier as most informative, being OmpR and NtrC the families with the highest TCS representation.

Osmolarity sensors EnvZ-OmpR and CpxA-CpxR (OmpR family) regulate the expression of outer membrane porins in Gram-negative bacteria. Porins control osmolar pressure in response to environmental changes, like from a free-living context to inside a host cell [22].

Gene *vicK* is part of *Bacillus subtilis* VicR-VicK system (also a member of OmpR family). It has been widely related to exopolysaccharide biosynthesis, biofilm formation and virulence factors expression in Gram-positives [23, 24]. Gene *vicK* is absent in an important group of non-pathogenic Firmicutes, including most non-pathogenic species of genus *Clostridium*. Seemingly, this feature allows the correct classification of these species and is also indicating a certain importance of the VicR-VicK system in some point of *Clostridium* pathogenesis.

The QseB-QseC system is involved in regulation of motility proteins [25], which are key virulence factors of many bacterial pathogens. Often, this system has pleiotropic effects over phenotypes including chemotaxis, adherence, host cell invasion, colonization and innate immune signaling [26]. It was identified in most distinctive pathogenic members of Gammaproteobacteria, including *Salmonella*, *Escherichia*, *Vibrio*, and *Shigella*. Surprisingly, it was absent in *Yersinia pestis* genomes.

Genes representing 5 TCS for NtrC family were selected. Among them we found PilS-PilR, another TCS involved in adherence and cell

invasion. This system is essential for type IV secretion systems induction in Neisseriaceae species, like *Kingella kingae* an increasingly common cause of septic arthritis, bacteremia, and osteomyelitis in young children [27]. Interestingly, orthologous genes of *pilR* were found in a small group of Gammaproteobacteria, including *Pseudomonas aeruginosa*, *Acitnetobacter baumanii* and *Legionella pneumophila*.

Toxins. Pathogenic bacteria have been developing a variety of strategies to manipulate host cell functions, often involving toxins [12]. These proteins have a wide range of action, causing different effects, like host cells deregulation, protein synthesis interruption or membrane damage [28–30]. A total of 76 different bacterial toxins were included in this work. Feature selection analysis selected 11 toxins for the model.

Streptolysin O (SLO) is a thiol-activated cytolysin, the effect of this pore-forming toxin is more subtle than simple lysis of host cells, and may include interference with immune cell function [31]. SLO is synthesized by more than 20 species of Gram-positive bacteria [32], and it is intimately involved in pathogenesis of *Arcanobacterium pyogenes*, *Clostridium perfringens*, *Listeria monocytogenes* and *Streptococcus pneumoniae* [31]. In this work, SLO was identified in pathogenic Firmicutes and absent in non-pathogenic species of this group. This gene is present in most pathogenic strains of *S. pyogenes*, *S. pneumoniae* and those species described by Billington et al. [31], but it is also present in pathogenic *Bacillus cereus*, *Streptococcus dysgalactiae* and *Gardnerella vaginalis*, the latter belonging to Actinobacteria.

Hemolysin II and thermolabile hemolysin are also pore-forming toxins selected by the model. The first is produced by pathogenic species of genus *Bacillus*, [33, 34] although, in this work, genes extremely similar to hemolysin II were also identified in all pathogenic strains of *Staphylococcus aureus*. Thermolabile hemolysin is characteristic of *Vibrio* species [35] as confirmed by the identification of this gene exclusively in *V. cholerae* and *V. vulnificus* strains.

Cytolethal distending toxin is able to block the host cell cycle between G2 and mitosis [28]. As described in previous works it was identified in a broad range of pathogenic bacteria including *Campylobacter* spp., *Salmonella enterica*, *Haemophilus ducreyi* and *Actinobacil-*

lus actinomycetemcomitans [31]. A/B toxins have similar effects in cell-cycle deregulation, affecting migration, morphogenesis, cell division [36] and membrane trafficking [37]. These were identified in *Clostridium difficile* and in many pathogenic strains of *Escherichia coli*, including O157:H7, O55:H7, O127:H6 and O103:H2. In addition to the contribution for classification, the presence of A/B toxin in these phylogenetically distant groups of possibly indicates horizontal gene transfer events between them.

Table 1.4: Classification performance for Group I and Group II.

	Taxon	Correctly classified	Wrongly classified	Accuracy
Group I	Chlamydiae	14	0	100%
	Deltaproteobacteria	26	0	100%
	Planctomycetes	3	0	100%
	Deinococcus-Thermus	3	0	100%
	Acidobacteria	3	0	100%
	Deltaproteobacteria	4	1	80%
	Chloroflexi	8	0	100%
	Cyanobacteria	27	1	96.4%
	Thermotogae	9	0	100%
	Other bacteria	19	0	100%
Group II	Actinobacteria	26	4	87%
	Alphaproteobacteria	24	2	92%
	Bacteroidetes	13	0	100%
	Betaproteobacteria	22	2	91%
	Deltaproteobacteria	5	0	100%
	Epsilonproteobacteria	8	1	89%
	Firmicutes	42	4	91%
	Gammaproteobacteria	38	4	90.5%
	Chloroflexi	6	0	100%
	Cyanobacteria	11	1	91%
	Deinococcus-Thermus	7	0	100%
	Other bacteria	13	0	100%

LPS biosynthesis. Lipopolysaccharides (LPS) are major components of the outer membrane of Gram-negative bacteria which can be recognized by the hostÕs toll-like receptor 4 (involved in inflammatory response). High concentrations of LPS can induce fever, increase heart rate, and lead to septic shock and death [38]. The model selected six (*lpxK*, *wapR*, *rgpA*, *gmhB*, *rfe* and *rfbP*) out of 94 genes, which code for proteins comprising different steps of typical Gram-negative LPS

biosynthesis. Tetraacyldisaccharide 4'-kinase (*lpxK*) catalyzes one of the last steps for Lipid A biosynthesis [39]. Genes *wapR* and *rgpA* produce rhamnosyltransferases, which add rhamnose to the polysaccharide backbone. In particular cases the incorporation of L- or R-rhamnose determines different glycoforms of the core region, leading to LPS variability, hence virulence [40]. Two genes are involved in O-antigen biosynthesis: *rfbP* codes for a glycosyltransferase responsible for the first step in O-antigen biosynthesis [41], while *rfe* (*wecA*) catalyzes the first membrane step of O-antigen and enterobacterial common antigen biosynthesis in *E. coli*. Its involvement in the virulence of Gram-negative bacteria has also been reported [42].

In spite of being selected by the model as relevant for classification, none of these genes showed a clear presence/absence pattern among pathogenic and non-pathogenic species. However, this does not mean they are not informative; on the contrary, these genes may be contributing to classification by an additive effect, being their individual inputs restricted to more particular groups.

Flagellar assembly and motility. Bacterial motility is a major factor in pathogenesis. This feature is involved in processes like biofilm formation, host cell colonization and bacterial spread inside the host [43]. Flagellar macromolecular machinery is the paradigm of bacterial motility, being present in a wide range of human pathogens, including *E. coli*, *S. enterica* and *P. aeruginosa* [44–46]. In the present work, 34 different genes involved in flagellum formation were investigated. Additionally, other 137 genes involved in different mechanisms related to bacterial motility (fimbrial proteins, adhesins, chemosensory proteins and regulatory proteins) were included.

Five genes directly involved in flagellar biosynthesis (*fliA*, *fliD*, *fliK*, *fliL* and *fliW*) were selected by the model. Gene *fliA* codes for σ^{28}, responsible for the regulation of flagellin biosynthesis. Inactivation experiments of *fliA* in *P. aeruginosa* cause non-motility, due to inability of expressing the flagellin gene [47]. The *fliD* gene codes for a structural component of the flagellar cap, which is important in host cell adhesion and colonization [48]. Gene *fliL* is dispensable for swimming in pathogenic species like *E. coli* and *S. enterica* [49], but it is essential for swarming (flagellar-dependent motility in solid medium) in

these species. Gene *fliK* is responsible for controlling flagellar hook length, which directly affects the performance of the flagella in producing translational motion [50]. Gene *fliW* codes for a new flagellin assembly protein in *Treponema pallidum* which has orthologous in many related species [51].

Table 1.5: Summary of the biological relevance for pathogenicity of a reduced subset of the selected 120 genes. The functional categories are described in Methods section.

Functional category	Genes	Comment
ABC	*sitC*, *hrtB*, *btuD*, *gluD*	Strong association between pathogens and the presence of transporters for metallic cations, vitamin B12, phosphate and amino acids
TCS&CH	*vicK*, *qseC*	VicK absent in most non-pathogenic Firmicutes. QseC is present in most pathogenic Gammaproteobacteria, but absent in Yersinia
LPS	*lpxK*, *wapR*, *rgpA*, *rfbP*	Genes involved in LPS biosynthesis did not show differences in presence/absence patterns between pathogens and non-pathogens
FLA&MOT	*flbP*, *fimH*, *fimI*, *pilA*	FlbP is found in pathogenic Spirochaetes. FimH and FimI are found in Enterobacteraceae. PilA is present in pathogens of a group of families inside Gammaproteobacteria
SS	*tatA*, *yscC*, *ppkA*	TatA is found in pathogenic Epsilonproteobacteria. YscC is part of T3SS from *Y. pestis* and many other pathogens. PpkA is part of T6SS from *Pseudomonas*
TOX	*slo*, *tlh*, *cdtC*	SLO is present in more than 20 pathogenic Gram-positive bacteria, including Firmicutes. Thermolabile hemolysin is exclusive from Vibrio. CdtC is present in a wide broad of pathogens including *Campylobacter*

Gene *flbB* is part of the flagellar motor exclusively in Spirochaetes [52]. In this work, this gene was found in pathogenic Spirochaetes and was absent in many other genomes, suggesting its importance for the correct classification of this group. Nevertheless, *flbB* homologues were also found in *Thermoanaerobacter* (Firmicutes). Independently of its role in the classification of pathogens, this finding questions the evolutionary origin of this flagellar motor, apparently exclusive for Spirochaetes.

Bacterial motility and host-cell adhesion are intimately related processes. Fimbria (type I pili) are filamentous proteinaceous surface appendages present in many Gram-negative bacteria [53, 54] that aid the adhesion process. In *E. coli*, fimbria are made of a repeating monomer, FimA, encoded by *fimA*. This gene is almost exclusively present in pathogenic Gammaproteobacteria and Betaproteobacteria, like *Escherichia*, *Salmonella*, *Acinetobacter* and *Burkhordelia*. FimH protein (encoded by *fimH*) is the most common adhesin located on the tip of type I fimbriae [55, 56]. Its expression, hence pilus formation, is regulated by gene *fimI*, which is essential for fimbriated phenotype. Specific mutations in *fimI* lead to pilus-negative phenotype in *E. coli* and *S. enterica* [57]. Both genes, *fimH* and *fimI*, were found exactly in the same group of species belonging to Enterobacteraceae family: *Salmonella*, *Escherichia*, *Proteus*, *Shigella* and *Klebsiella*. This supports the functional relationship of both genes and also denotes the importance of them for classification of this family of pathogenic Gammaproteobacteria.

Another relevant pili aparatus is the type IV system. This macromolecular machinery is present in Gram-negative bacteria and in at least one Gram-positive [58]. Type IV pili are highly pleiotropic, being involved in bacterial motility, adhesion, immune escape, biofilm formation, secretion and phage transduction. The most relevant selected gene for this pili system was *pilA*, which codes for pilin, the major component of filament. It is present in most pathogenic *Clostridium* (*C. perfringes*, *C. tetani*, *C. difficile* and *C. botullinum*). PilA is also present in pathogenic members of a group of families belonging to Gammaproteobacteria (Vibrionaceae, Pseudomonadaceae, Francisellaceae, Moraxellaceae). Interestingly, *pilA* is absent in pathogenic Enterobacteraceae, so the combination of three genes (*pilA*, *fimH* and *fimI*) seems to explain the discrimination of most pathogenic Gammaproteobacteria with respect to the rest of nonpathogenic bacteria and even distinguishing between two enormous phyolgenetic groups inside this taxon.

Secretion systems. Several differences in secretion systems exist between Gram-positive and Gram-negative bacteria. Protein secretion across the inner membrane of both kinds of organisms generally involves the same Sec-dependent pathway, although other routes have

been identified, i.e. Twin-arginine translocation (Tat) [59–61]. Translocation across Gram-negatives inner membrane results in release of products into the periplasmic space. Hence, these bacteria have developed several types of secretion systems which carry molecules from the periplasmic space to the cell surface or extracellular matrix. These secretory pathways of Gram-negatives can be classified into six different groups: type I to VI secretion systems (T1SS-T6SS). The presence/absence of 73 different genes coding for both shared secretory pathways (like Sec or Tat) and for T1SS-T6SS was tested. The model selected 13 genes as the most relevant to explain class differences.

Genes for Sec system were not selected by the model. For Tat system the tatA gene was selected; it codes the major pore-forming subunit for translocation complex [62]. Homologues of *tatA* have been identified in a wide range of human pathogens, including *E. coli* O:157, *Vibrio cholerae*, *Mycobacterium tuberculosis*, *Listeria monocytogenes* and *Staphylococcus aureus* [63]. Moreover, this gene has orthologous in all Epsilonproteobacteria analyzed in this work, except for the non-pathogenic *Sulfurovum* sp. NBC37-1. Even though *tatA* was selected as an important feature for classification, a clear presence/absence pattern between pathogenic and non-pathogenic species was not observed.

Gene *yscC* encodes a key protein of the archetypical T3SS of *Yersinia pestis*, the infective agent of human plague. YscC orthologs are now identified in more than a dozen of pathogens [64], including *Salmonella enterica*, *Shigella flexneri* [65] and enteropathogenic *E. coli* [66]. Beyond these well-known examples, we identified the presence of *yscC* orthologs only in species belonging to Gammaproteobacteria and Betaproteobacteria, being absent in a great number of non-pathogenic species.

T4SS have been described in several organisms including *Bordetella pertussis* [67], *Legionella pneumophila* [68], *Brucella suis* [69], *Bartonella henselae* [59], and *Helicobacter pylori* [70]. VirB2, coded by *virB2*, is major component of T4SS pilus and has an important role in secretion [71]. Beyond its identification in the species mentioned above, *virB2* is present in some genomes of well-known pathogens with different taxonomic context: *Campylobacter jejuni* subsp. *jejuni* 81-176 (Epsilonproteobacteria), *Klebsiella pneumoniae* subsp. *pneumoniae* NTUH-K2044 (Gammaproteobacteria), *Neorickettsia sennetsu* str. Miyayama

(Alphaproteobacteria) and three *Burkholderia* sp. species (Betaproteobacteria). This suggests an important role of T4SS in pathogenic processes, even in species with different pathogenic mechanisms.

T6SS have been found in species from a wide taxonomic range [72], comprising most bacterial groups included in this work. Two T6SS genes were selected: *ppkA* codes for a serine/threonineprotein kinase that phosphorylates protein FHA (encoded by *fha1*). The phosphorylation initiates a signal transduction cascade that results in T6SS assembly and function. Mutation of *P. aeruginosa fha1* gene resulted in defective secretion of Hcp1, an essential protein for pathogenesis as demonstrated by attenuated virulence phenotype observed in vivo [73]. Both *fha1* and *ppka* were identified in *P. fluorescens* and *P. mendocina* and all strains of *P. aeruginosa*. Interestingly, the absence of these genes in other genomes shows the great importance of their presence for the classification of these organisms exclusively. Moreover, the high correlation in the presence of both genes in the same genomes evidences their functional relationship.

1.3.4 Phylogenetic distribution of virulence genes. In the sections above we discused the biological meaning of some genes selected by the model, emphasizing their presence/absence patterns among pathogens and non-pathogens and their importance in the development of pathogenic phenotypes. Here we give an integrative overview of virulence genes distribution along bacterial phylogeny, taking into account their frequency bias among pathogenic and non-pathogenic organisms. Fisher exact test (p-value <0.001) was used to select genes with significant differences in their presence/absence patterns for each functional category inside each taxonomic group. Then, gene frequency was calculated among pathogens and non-pathogens for those selected genes, separated by functional category. Finally, individual genes frequencies were added inside each group and normalized over the total number of genes belonging to each functional category.

Fig. 1.5 shows normalized frequency values for genes belonging to each functional category, taking into account the phylogenetic relationships between studied taxonomic groups. Some expected patterns arise from these results, for example toxins are exclusively overrepresented in pathogenic species. This is expectable taking into account

the biological purpose of toxins; it would be highly improbable that pathogenicity in a certain species was determined by the absence of a toxin that is present in the non-pathogenic species of the group. ABC transporters seem to be the most variable functional category along the phylogeny, it is positive (associated to pathogenic organisms) in Gammaproteobacteria, Betaproteobacteria and Firmicutes, and negative (associated to nonpathogenic organisms) in Alphaproteobacteria and Actinobacteria. This is coherent with the wide range of functions that ABC transporters can perform; for example the presence of amino acid importers can be essential for pathogenesis of species that have lost biosynthetic genes, however, it is not contradictory with the presence of these kind of transporters in non-pathogenic species.

The most powerful association between pathogens and high gene frequencies is observed in Gammaproteobacteria, evidencing the importance of these kinds of genes for pathogenic species of this group, which is mainly composed of enteropathogens. The most striking result of this analysis is the pattern observed for Alphaproteobacteria, totally opposite to the phylogenetically related Gammaproteobacteria. The first question that rises is why genes previously thought of as mostly present in pathogenic species, are highly frequent in non-pathogenic species of this taxon. Marine environments contain the major component of non-pathogenic Alphaproteobacteria biodiversity. A recent study [74] showed that out of 119 marine bacteria, 60 had homologues to known virulence genes from pathogenic bacteria. Interestingly, new insights in host-pathogen interactions propose a wider ecological and evolutionary perspective to better understanding the life strategy of pathogenic bacteria [75], suggesting that functions have evolved over a long time in nature and then recruited through horizontal gene transfer to perform similar or different functions in more recently emerging pathogenic species. This hypothesis opens a three-step way of thinking about how natural selection plays a role in the emergence of bacterial pathogens. First, the random appearance and fixation of new genes in bacteria colonizing inaccessible environments generate a reservoir of species carrying potentially virulent genes. Second, these bacteria can contact human hosts by movement through intermediate hosts in which they live as commensals or they can transfer virulent genes horizontally to other human-adapted bacteria. Third,

positive selection over the most successful species determines the fixation of virulence genes that let bacteria to damage or survive inside human cells. The high frequency of virulence-related genes in nonpathogenic Alphaproteobacteria might be explained by the emergence of these kinds of genes in common ancestors for Gammaproteobacteria and Alphaproteobacteria. Then, the branch that originated Alphaproteobacteria conserved these genes in both pathogenic and nonpathogenic species. In contrast, Gammaproteobacteria could have acquired these functions by horizontal gene transfer, to produce the actual scenario of high frequency in pathogenic species and low frequency in nonpathogenic ones.

Two groups (Spirochaetes and Epsilonproteobacteria) showed very few genes with significant differences according to Fisher exact test. This reveals that for these two taxonomic groups there are no clear presence/absence patterns among genes of pathogenic and nonpathogenic species but, in spite of this, our model is able to assign each organisms to the correct class with high accuracy. This is particularly interesting because our model is using information coded in high-dimensional spaces, leaving behind the simple presence/absence patterns. Moreover, here we could identify only some particular associations between phylogeny topology and functional categories, suggesting that, in general, the functional importance of these genes varies along bacterial taxonomy. The lack of general patterns between the presence of functional categories and phylogenetically related groups supports the notion that most virulence-related genes are spread among bacteria by horizontal gene transfer. Probably our method is taking benefit of this scenario, being able to correctly classify organisms independently of their taxonomic context, based on widely spread genes along bacterial phylogeny.

1.3.5 Misclassified organisms. A group of 28 out of the 648 genomes tested were systematically misclassified by the model. We defined a genome to be misclassified if it was assigned to the wrong class, at least in 50% of 20 consecutive classifications (Supp. Tab. 1.7). Ten out of these 28 are labeled as human pathogens but the model returned them as non-pathogenic, while 18 out of 28 are labeled as nonpathogenic but were classified as human pathogens. Most cases of mis-

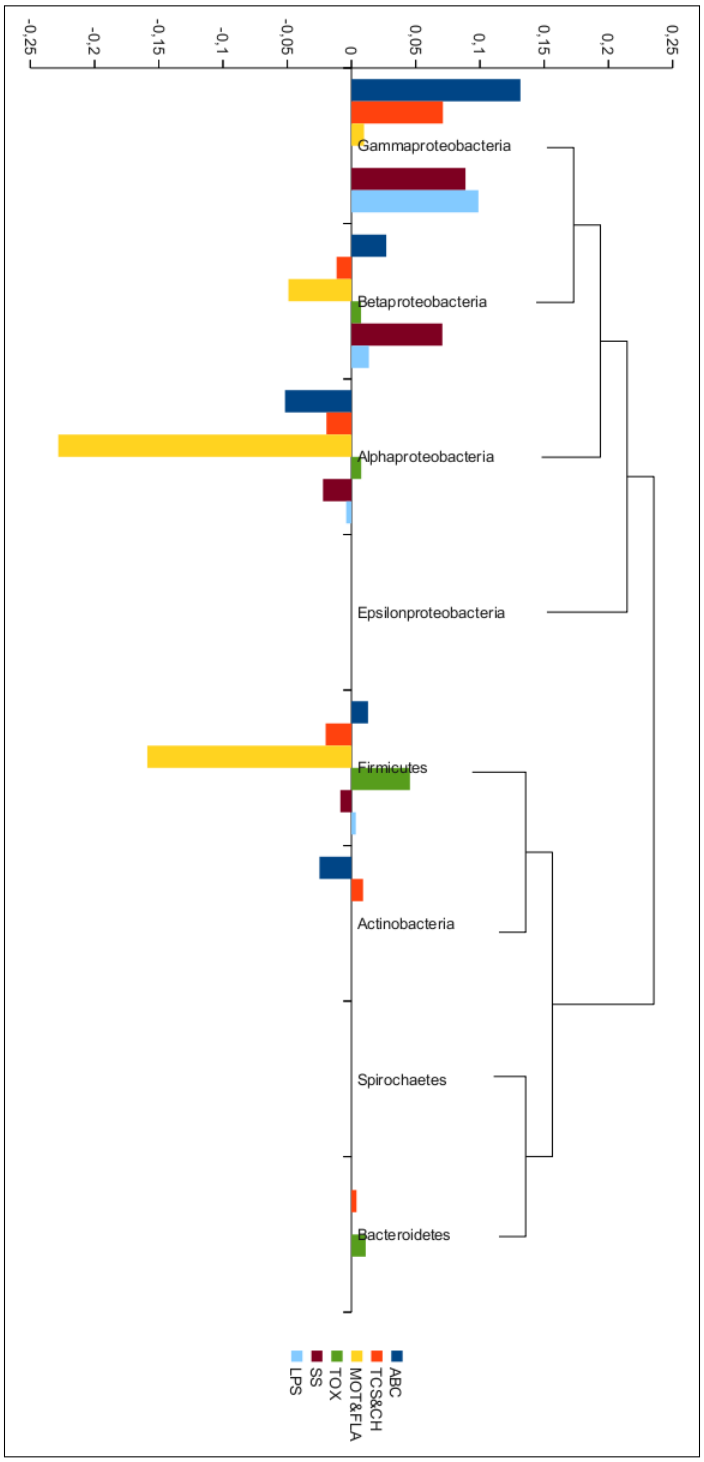


Figure 1.5: Phylogenetic distribution of virulence genes. Each functional category of virulence-related genes is represented as a vertical bar. Positive values denote association of a particular functional category with pathogenic species of a certain taxonomic group, while negative values with non-pathogenic species. Taxa are grouped according to phylogenetic relationships. In graph legend: ABC: ABC transporters, TCS&CH: two-component systems and chemotaxis, MOT&FLA: motility and flagellar assembly, TOX: toxins, SS: secretion systems, LPS: LPS biosynthesis.

classification are observed in species with a big number of sequenced genomes of different strains. This is the case of *Staphylococcus aureus*, an important human pathogen. Thirteen out of the 14 genomes of different strains of this species were well classified as human pathogens. Nevertheless, the strain *S. aureus* subsp.*aureus* MRSA252 was assigned to the non-pathogenic class. Comparison of present/absent genes for all *S. aureus* genomes showed that gene *hlyII* (coding for hemolysin II) was absent in *S. aureus* subsp. *aureus* MRSA252 while present in the rest. This was the only difference between these genomes; moreover gene *hlyII* was one of the 11 toxin-coding genes selected as more informative during the feature selection process. On the one hand, this fact shows that for a particular species even the presence of a single feature is determining the classification of the genome as pathogenic or non-pathogenic, indicating a great power of some genes in determining the class assignment by the model. On the other hand, it is possible to misclassify genomes due to a particular gene loss, especially in those cases of high genetic variability among strains of certain species.

For misclassified genomes that do not have other well-classified strains belonging to the same species, it is not possible to assess the present/absent comparison to find differences in gene patterns. In these cases, misclassification can be explained by inherent errors of SVM model construction or because the features (groups of orthologous genes) originally used to determine the presence/absence matrix, might not be informative enough to reach a 100% classification performance. However, in some cases it is possible to propose a biological explanation for misclassification, based on the particular ecological and genetic features of some species.

The first example is *Bordetella petrii* (Betaproteobacteria) which is originally labeled as non-pathogenic, but the model classifies it as pathogenic. This could be primarily seen as a classification error, but there is strong evidence that supports this species is an emerging human pathogen. Though being an environmental isolate, the sequenced *B. petrii* DSM12804 strain also encodes proteins related to virulence factors of the pathogenic Bordetellae, including the filamentous hemagglutinin, which is a major colonization factor of *B. pertussis*. The genomic analysis of *B. petrii* suggests an evolutionary link between free-living environmental bacteria and the host-restricted obli-

gate pathogenic Bordetellae [76]. Moreover, clinical isolates of *B. petrii* have been recently described to cause, for example, mandibular osteomyelitis [77] or supurative mastoiditis [78].

Other example comprises a group of 6 marine non-pathogenic Alphaproteobacteria (*Rhodobacter capsulatus*, *Erythrobacter litoralis*, *Rhodopseudomonas palustris*, *Novosphingobium aromaticivorans*, *Parvularcula bermudensis* and *Sphingobium japonicum*), wrongly classified as pathogenic. As explained in the section above, Alphaproteobacteria have the highest frequency of virulence-related genes in non-pathogenic species. The 6 misclassified species shared the presence of 9 genes involved in secretion processes, supporting the findings of Persson *et al.* [74] regarding the extensive appearance of these kinds of genes in marine bacteria. Despite this, only 6 out of 88 Alphaproteobacteria were misclassified, indicating that the classification model can deal with unexpectedly biased gene frequencies towards non-pathogenic organisms without compromising classification performance.

1.3.6 Model sensitivity. A simple approach to evaluate the sensitivity of the constructed model is to assess the propensity of label shift (pathogens to non-pathogens and vice versa). This experiment was implemented for each taxonomic group in the dataset by artificially modifying presence/absence vectors. For each genome those present genes were systematically "turned off" one at a time, running the classification model each time and recording in which cases a category shift occurred. The same strategy was used to "turn on" those genes which were originally absent.

The change from non-pathogen to pathogen was lead by a group of 14 genes, which were mainly toxin-coding genes (5) and TCS (5). These two functional categories together comprise 23 of the genes that influence the category shifting in the mentioned direction, evidencing a great importance of these features as exclusive determinants of bacterial pathogenicity. Individually, the presence of any of these genes is able to change a number of organisms ranging from 78 to 153, depending on the gene. The most extreme is the case of SLO toxin, whose presence determines that 153 species change from non-pathogens to pathogens.

Changing from pathogen to non-pathogen is mainly determined by gene "turn off". A group of 9 genes are responsible for category shifting in this direction, changing the classification of 10 to 96 species. It is worth mentioning that the gene coding for the SLO toxin is one of the most influential; this makes sense, since the gain of this gene provoked a label change to pathogen, it is expectable that losing it defines a label change to non-pathogen.

1.3.7 Software development: the BacFier. BacFier v1.0 was implemented as a Java software, and hence platform independent, in order to make it easier for the common user to work with the model. A simple interface allows the user to upload the genome sequence (finished or unfinished) of the organism of interest. The genome is used as query to perform BLAST against the final set of 120 orthologous groups (selected as explained in section "Model construction") creating a presence/absence vector for the genome. The vector is evaluated with a SVM model, and an outcome (pathogen/non-pathogen) is produced associated to a probability.

Moreover, the sensitivity analysis described in the previous section can be automatically performed with the software, this is assessed by selectively "turning off" or "turning on" desired genes in the presence/absence vector and re classifying the result. This might indicate genes that are likely to change the label of the organism, so that one can pay more attention to them and corroborate their status of presence/absence. Furthermore, this strategy becomes crucial when inputing an unfinished genome. In this situation, the absence of some genes important for pathogenicity could be determined by the unfinished status of the genome, so if prediction result is non-pathogenic, the user can systematically "turn on" those absent genes until the model shift to pathogenic. Then, the real presence of genes that determined the shift can be investigated by a more refined search or by other methods, like PCR. BacFier v1.0 is freely available under `https://code.google.com/archive/p/bacfier/`.

1.3.8 Conclusions. The constructed SVM model classifies bacterial genomes in human pathogens and non pathogens with 95.4% of average accuracy. To the best of our knowledge, this is the statistical model

with this purpose that achieves the highest accuracy reported so far. Moreover, our method classifies bacterial genomes independently of their taxonomic context, in contrast to other similar approaches that only take into account a certain part of bacterial diversity, being useful only to classify specific taxa [16]. Our statistical learning approach is grounded on the biological meaning of the selected genes and supported by the fact that bacterial pathogenicity can be explained by the presence or absence of a set of specific genes that code for virulence determinants. The application of BacFier v1.0 may be useful for clinical or industrial purposes, for example to determine if a new sequenced strain could be pathogenic for humans.

1.4 Methods

1.4.1 Data selection and matrix construction. Complete genome sequences from all available bacteria were downloaded from the National Center for Biotechnology Information (NCBI). Over 1000 genomes were obtained and from those organisms, we originally kept 848 that were labeled as human pathogens or non-pathogens. This set of bacteria comprehends 22 taxonomic groups. In this work, we focused only on human pathogens; if a certain species was a multi-host pathogen including humans, it was considered human pathogen. By the contrary, if a certain species was a multi-host pathogen or a pathogen of other host different from human, it was excluded from the dataset considered.

Eight gene functional categories that we considered related to pathogenicity were determined. These are toxins, chemotaxis proteins, ABC transporters, motility proteins, LPS biosynthesis, two-component systems, flagellar assembly and secretion systems. Orthologous groups from proteins coded by genes belonging to these categories were downloaded from KEGG Orthology database (`http://www.genome.jp/kegg/ko`), all the categories together resulted in 814 orthologous groups. With this data, we built a presence/absence table showing which orthologous groups (genes/proteins) were present or absent in the organisms considered. We selected local protein BLAST [79] searches to perform orthologous genes determination. Not only does this approach absolve us from using a refined orthologous search

method (which can be much more laborious and time-consuming), but it also provides good enough accuracy in orthologous determination. In this case, our method must be robust and tolerant enough to identify possible false positive or false negative orthologs.

BLAST searches were performed formatting the 814 ortbologous groups and querying the organisms. If an alignment between an organism and a gene (member of an orthologous group) was "good enough" (see below), then we considered the gene (orthologous group) as present in the organism, otherwise as absent. This, is represented as a binary (0/1) matrix with organisms as rows and orthologous groups as columns. We defined "good" alignments as the ones having a percentage of identity higher than 90%, length of the alignment larger than 90% of the gene's length and an e-value <0.001. Further analyses were made on 648 genomes belonging to 8 of the 22 taxonomic groups: Actinobacteria, Alphaproteobacteria, Bacteroidetes/Chlorobi, Betaproteobacteria, Epsilonproteobacteria, Firmicutes, Gammaproteobacteria and Spirochaetes, since there were not enough genes available for the other groups. However, these excluded genomes were then used as part of external groups to further test the constructed model.

1.4.2 Model construction. In this work a machine learning approach based on a cross-fold validation with in-fold feature selection was developed. This technique ensures that particular predictions are not biased by over-selected features or over-fitting since each prediction is performed without using the sample in neither the feature selection nor the classifier building process. Algorithm 1 shows the methodology.

The number of folds (nfold) was set to 10 and the feature selection routine was SVMAttributeEval from Weka [80]. Regarding the classification algorithm, a Support Vector Machine (SVM) was employed. The SVM method performs the classification by constructing an N-dimensional hyperplane that optimally separates the data into two classes. In this case classes are labeled as human pathogens and non-pathogens. The raw dataset of variables is defined by the presence/absence of orthologous groups in the genomes of the organisms considered. It is important to note that the taxonomy is not used as another

Algorithm 1 Pseudocode of cross-fold validation with in-fold feature selection (be X = whole set of samples).

```
1: for i ← 1 →nfolds do
2:     Define validation set VS ← samples in i
3:     Define training set TS ← X - VS
4:     Perform feature selection over TS samples
5:     Train classifier using TS
6:     Perform prediction of VS samples with previous classifier
7: end for
```

variable in the model since it would introduce an artificial separation in the SVM model training.

Following the spirit of Occam's razor, in this work a linear SVM model is proposed. Although the number of genes looks relatively large, it is worth to mention that the model variables encode low level information related to gene presence/absence in each organism. Also, it is well known that linear SVM models benefit from using these kinds of variables since higher dimensions allow easier class separation. The subroutine libsvm in Weka was also employed [80].

A final analysis was done in order to determine an aproppiate number of features to retain. Experiments were carried out considering 30, 60, 90, 120, 150, 200 and 841 (entire set of genes) features. The accuracy obtained in each case was 90%, 93.5%, 94.4%, 95.4%, 95.5%, 94.9% and 92.1% respectively. A set of 120 genes was then considered, as they represent a reasonable tradeoff between accuracy prediction and the number of genes used for prediction.

From Alg. 1 is clear that a different set of features can be selected in each loop of the cross-validation procedure. However, it is necessary to find a final set of genes to build a classification model and check and external validation set (for practical purpose) or predict pathogenicity of new sequenced bacteria. A common solution is to employ a voting scheme that sums how many times a feature is selected in each loop of Algorithm 1. In this particular case, the list of genes selected is available in Supp. Tab. 1.6.

1.4.3 Y-randomization test. Since in this work a binary occurrence matrix is used to represent the presence/absence of genes in a set of

organisms, the number of calculated variables is high, as expected. In this particular case, the number of genes is 814. A feature selection technique further reduced the set to the 120 most significant variables. Although this meets the rule of thumb that states the ratio between number of samples (648 organisms) and variables (120) must be greater than 5 [81], problems associated with chance correlation could still arise. This is a major concern when the prediction model is expected to be reliable in terms of generalizability.

The y-randomization validation method tries to observe the influence of chance when fitting any given data. This is done by deliberately destroying the relationship between the target y and the independent variables x (genes, in this case). This is done by randomly shuffling the y data, preserving all x data untouched, and retraining the learning algorithm. A common pitfall is to apply the y-randomization procedure but using the same set of variables resulting from the feature selection process. Following the good-practice procedures, in this work the test was carried out using the full set of variables, so there was no "overestimation" (in the sense of chance correlation).

In this work we have two classes, so the expected behavior was to obtain an accuracy of roughly 50% in the y-randomization test (since 50% is the probability of a "good" prediction when no relation is found between variables and targets, the same as a random assignment of predicted labels). In this work the y-randomization procedure was carried out 100 times (Supp. Fig. 1.1).

1.4.4 Genes significance and frequency calculation. In order to weight the importance of each functional category for each taxonomic group, we selected those genes with statistically significant presence/absence patterns inside pathogens and non-pathogens. Fisher exact test was applied to genes belonging to each functional category for each taxonomic group. Those genes with p-value <0.001 were taken into account. Then, the frequency of those genes was calculated for pathogenic and non-pathogenic species of each taxonomic group, as the number of presences over the total number of organisms inside the group. Finally, for a certain functional category, the significance value was calculated as the accumulated frequency of those genes significant to the category, and normalized over the total number of genes belong-

ing to it. For a better graphical visualization of Fig. 1.5, frequencies in non-pathogenic organisms were multiplied by -1, in this way positive values are associated with pathogenic organisms while negative with non-pathogenic ones.

1.5 Acknowledgments

We thank Natalia Rego for thorough reading and insightful comments on the manuscript.

1.6 References

[1] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J.-F. Tomb, B. A. Dougherty, J. M. Merrick, et al., *Science* **1995**, *269*, 496–512.

[2] C. M. Fraser, J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton, R. D. Fleischmann, C. J. Bult, A. R. Kerlavage, G. Sutton, J. M. Kelley, et al., *Science* **1995**, *270*, 397–404.

[3] J.-F. Tomb, O. White, A. R. Kerlavage, R. A. Clayton, G. G. Sutton, R. D. Fleischmann, K. A. Ketchum, H. P. Klenk, S. Gill, B. A. Dougherty, et al., *Nature* **1997**, *388*, 539–547.

[4] M. Kuroda, T. Ohta, I. Uchiyama, T. Baba, H. Yuzawa, I. Kobayashi, L. Cui, A. Oguchi, K.-i. Aoki, Y. Nagai, et al., *The Lancet* **2001**, *357*, 1225–1240.

[5] N. T. Perna, G. Plunkett, V. Burland, B. Mau, J. D. Glasner, D. J. Rose, G. F. Mayhew, P. S. Evans, J. Gregor, H. A. Kirkpatrick, et al., *Nature* **2001**, *409*, 529–533.

[6] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, et al., *Nature* **2005**, *437*, 376–380.

[7] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, et al., *nature* **2008**, *456*, 53–59.

[8] E. V. Sokurenko, V. Chesnokova, D. E. Dykhuizen, I. Ofek, X.-R. Wu, K. A. Krogfelt, C. Struve, M. A. Schembri, D. L. Hasty, *Proceedings of the National Academy of Sciences* **1998**, *95*, 8922–8926.

[9] G. M. Conenello, D. Zamarin, L. A. Perrone, T. Tumpey, P. Palese, *PLoS Pathog* **2007**, *3*, e141.

[10] H. Marjuki, C. Scholtissek, J. Franks, N. J. Negovetich, J. R. Aldridge, R. Salomon, D. Finkelstein, R. G. Webster, *Archives of virology* **2010**, *155*, 925–934.

[11] L. Spangenberg, F. Battke, M. Graña, K. Nieselt, H. Naya, *Bioinformatics* **2011**, *27*, 2782–2789.

[12] E. Oswald, J.-P. Nougayrède, F. Taieb, M. Sugai, *Current opinion in microbiology* **2005**, *8*, 83–91.

[13] J. A. Lanie, W.-L. Ng, K. M. Kazmierczak, T. M. Andrzejewski, T. M. Davidsen, K. J. Wayne, H. Tettelin, J. I. Glass, M. E. Winkler, *Journal of bacteriology* **2007**, *189*, 38–51.

[14] T. Baba, T. Bae, O. Schneewind, F. Takeuchi, K. Hiramatsu, *Journal of bacteriology* **2008**, *190*, 300–310.

[15] S. J. H. Sui, A. Fedynak, W. W. Hsiao, M. G. Langille, F. S. Brinkman, *PloS one* **2009**, *4*, e8094.

[16] M. Andreatta, M. Nielsen, F. M. Aarestrup, O. Lund, *PloS one* **2010**, *5*, e13680.

[17] D. C. Rees, E. Johnson, O. Lewinson, *Nature reviews Molecular cell biology* **2009**, *10*, 218–227.

[18] L. Rohmer, D. Hocquet, S. I. Miller, *Trends in microbiology* **2011**, *19*, 341–348.

[19] S. A. West, A. Buckling, *Proceedings of the Royal Society of London B: Biological Sciences* **2003**, *270*, 37–44.

[20] N. A. Moran, *Cell* **2002**, *108*, 583–586.

[21] A. M. Stock, V. L. Robinson, P. N. Goudreau, *Annual review of biochemistry* **2000**, *69*, 183–215.

[22] N. J. Shikuma, F. H. Yildiz, *Journal of bacteriology* **2009**, *191*, 4082–4096.

[23] S. Dubrac, I. G. Boneca, O. Poupel, T. Msadek, *Journal of bacteriology* **2007**, *189*, 8257–8269.

[24] M. D. Senadheera, B. Guggenheim, G. A. Spatafora, Y.-C. C. Huang, J. Choi, D. C. Hung, J. S. Treglown, S. D. Goodman, R. P. Ellen, D. G. Cvitkovitch, *Journal of bacteriology* **2005**, *187*, 4064–4076.

[25] V. Sperandio, A. G. Torres, J. B. Kaper, *Molecular microbiology* **2002**, *43*, 809–821.

[26] C. Josenhans, S. Suerbaum, *International Journal of Medical Microbiology* **2002**, *291*, 605–614.

[27] T. E. Kehl-Fie, E. A. Porsch, S. E. Miller, J. W. StGeme, *Journal of bacteriology* **2009**, *191*, 4976–4986.

[28] C. A. Whitehouse, P. B. Balbo, E. C. Pesci, D. L. Cottle, P. M. Mirabito, C. L. Pickett, *Infection and immunity* **1998**, *66*, 1934–1940.

[29] O. Marchès, T. N. Ledger, M. Boury, M. Ohara, X. Tu, F. Goffaux, J. Mainil, I. Rosenshine, M. Sugai, J. De Rycke, et al., *Molecular microbiology* **2003**, *50*, 1553–1567.

[30] B. Gebert, W. Fischer, E. Weiss, R. Hoffmann, R. Haas, *Science* **2003**, *301*, 1099–1102.

[31] S. J. Billington, B. H. Jost, J. G. Songer, *FEMS microbiology letters* **2000**, *182*, 197–205.

[32] J. E. Alouf, *Pharmacology & therapeutics* **1980**, *11*, 661–717.

[33] M. Sinev, Z. Budarina, I. Gavrilenko, Tomashevskiĭ.

[34] Z. I. Budarina, M. A. Sinev, S. G. Mayorov, A. Y. Tomashevski, I. V. Shmelev, N. P. Kuzmin, *Archives of microbiology* **1994**, *161*, 252–257.

[35] X.-H. Zhang, B. Austin, *Journal of Applied Microbiology* **2005**, *98*, 1011–1019.

[36] A. B. Jaffe, A. Hall, *Annu. Rev. Cell Dev. Biol.* **2005**, *21*, 247–269.

[37] A. J. Ridley, *Traffic* **2001**, *2*, 303–310.

[38] M. Yamamoto, S. Sato, H. Hemmi, S. Uematsu, K. Hoshino, T. Kaisho, O. Takeuchi, K. Takeda, S. Akira, *Nature immunology* **2003**, *4*, 1144–1150.

[39] C. R. Raetz, Z. Guan, B. O. Ingram, D. A. Six, F. Song, X. Wang, J. Zhao, *Journal of lipid research* **2009**, *50*, S103–S108.

[40] K. K. Poon, E. L. Westman, E. Vinogradov, S. Jin, J. S. Lam, *Journal of bacteriology* **2008**, *190*, 1857–1865.

[41] G. L. Murray, S. R. Attridge, R. Morona, *Journal of bacteriology* **2006**, *188*, 2735–2739.

[42] B. Al-Dabbagh, D. Mengin-Lecreulx, A. Bouhss, *Journal of bacteriology* **2008**, *190*, 7141–7146.

[43] M. E. Hibbing, C. Fuqua, M. R. Parsek, S. B. Peterson, *Nature Reviews Microbiology* **2010**, *8*, 15–25.

[44] L. Yim, L. Betancor, A. Martínez, C. Bryant, D. Maskell, J. A. Chabalgoity, *Applied and environmental microbiology* **2011**, *77*, 7740–7748.

[45] G. A. O'Toole, R. Kolter, *Molecular microbiology* **1998**, *30*, 295–304.

[46] T. K. Wood, A. F. G. Barrios, M. Herzberg, J. Lee, *Applied microbiology and biotechnology* **2006**, *72*, 361–367.

[47] M. Starnbach, S Lory, *Molecular microbiology* **1992**, *6*, 459–469.

[48] A. Tasteyre, M.-C. Barc, A. Collignon, H. Boureau, T. Karjalainen, *Infection and immunity* **2001**, *69*, 7937–7940.

[49] G. J. Schoenhals, R. M. Macnab, *Microbiology* **1999**, *145*, 1769–1775.

[50] R. C. Waters, P. W. O'Toole, K. A. Ryan, *Protein Science* **2007**, *16*, 769–780.

[51] B. Titz, S. V. Rajagopala, C. Ester, R. Häuser, P. Uetz, *Journal of bacteriology* **2006**, *188*, 7700–7706.

[52] J. Liu, T. Lin, D. J. Botkin, E. McCrum, H. Winkler, S. J. Norris, *Journal of bacteriology* **2009**, *191*, 5026–5036.

[53] G. E. Soto, S. J. Hultgren, *Journal of bacteriology* **1999**, *181*, 1059–1071.

[54] P. Aprikian, G. Interlandi, B. A. Kidd, I. Le Trong, V. Tchesnokova, O. Yakovenko, M. J. Whitfield, E. Bullitt, R. E. Stenkamp, W. E. Thomas, et al., *PLoS Biol* **2011**, *9*, e1000617.

[55] C. H. Jones, J. S. Pinkner, R. Roth, J. Heuser, A. V. Nicholes, S. N. Abraham, S. J. Hultgren, *Proceedings of the National Academy of Sciences* **1995**, *92*, 2081–2085.

[56] E. Hahn, P. Wild, U. Hermanns, P. Sebbel, R. Glockshuber, M. Häner, N. Taschner, P. Burkhard, U. Aebi, S. A. Müller, *Journal of molecular biology* **2002**, *323*, 845–857.

[57] M. L. Valenski, S. L. Harris, P. A. Spears, J. R. Horton, P. E. Orndorff, *Journal of bacteriology* **2003**, *185*, 5007–5011.

[58] L. Craig, J. Li, *Current opinion in structural biology* **2008**, *18*, 267–277.

[59] R. Schulein, C. Dehio, *Molecular microbiology* **2002**, *46*, 1053–1067.

[60] H. Mori, K. Ito, *Trends in microbiology* **2001**, *9*, 494–500.

[61] C. Robinson, A. Bolhuis, *Nature Reviews Molecular Cell Biology* **2001**, *2*, 350–356.

[62] M. Müller, *Research in microbiology* **2005**, *156*, 131–136.

[63] K. Dilks, R. W. Rose, E. Hartmann, M. Pohlschröder, *Journal of bacteriology* **2003**, *185*, 1478–1483.

[64] G. R. Cornelis, F. Van Gijsegem, *Annual Reviews in Microbiology* **2000**, *54*, 735–774.

[65] P. J. Sansonetti, *American Journal of Physiology-Gastrointestinal and Liver Physiology* **2001**, *280*, G319–G323.

[66] J. Celli, W. Deng, B. B. Finlay, *Cellular microbiology* **2000**, *2*, 1–9.

[67] K. M. Farizo, T. Huang, D. L. Burns, *Infection and immunity* **2000**, *68*, 4049–4054.

[68] S. D. Zink, L. Pedersen, N. P. Cianciotto, Y. A. Kwaik, *Infection and immunity* **2002**, *70*, 1657–1663.

[69] M. L. Boschiroli, S. Ouahrani-Bettache, V. Foulongne, S. Michaux-Charachon, G. Bourg, A. Allardet-Servent, C. Cazevieille, J.-P. Lavigne, J. P. Liautard, M. Ramuz, et al., *Veterinary microbiology* **2002**, *90*, 341–348.

[70] S. Backert, Y. Churin, T. F. Meyer, *The Keio journal of medicine* **2002**, *51*, 6–14.

[71] G. Schröder, C. Dehio, *Trends in microbiology* **2005**, *13*, 336–342.

[72] L. E. Bingle, C. M. Bailey, M. J. Pallen, *Current opinion in microbiology* **2008**, *11*, 3–8.

[73] E. Potvin, D. E. Lehoux, I. Kukavica-Ibrulj, K. L. Richard, F. Sanschagrin, G. W. Lau, R. C. Levesque, *Environmental microbiology* **2003**, *5*, 1294–1308.

[74] O. P. Persson, J. Pinhassi, L. Riemann, B.-I. Marklund, M. Rhen, S. Normark, J. M. González, Å. Hagström, *Environmental microbiology* **2009**, *11*, 1348–1357.

[75] M. J. Pallen, B. W. Wren, *Nature* **2007**, *449*, 835–842.

[76] R. Gross, C. A. Guzman, M. Sebaihia, V. A. M. dos Santos, D. H. Pieper, R. Koebnik, M. Lechner, D. Bartels, J. Buhrmester, J. V. Choudhuri, et al., *BMC genomics* **2008**, *9*, 1.

[77] N. K. Fry, J. Duncan, H. Malnick, M. Warner, A. J. Smith, M. S. Jackson, A. Ayoub, *Emerg Infect Dis* **2005**, *11*, 1131–3.

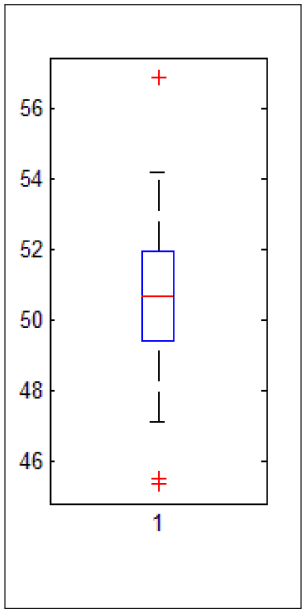
[78] D Stark, L. Riley, J Harkness, D Marriott, *Journal of medical microbiology* **2007**, *56*, 435–437.

[79] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **1990**, *215*, 403–410.

[80] E. Frank, M. Hall, L. Trigg, G. Holmes, I. H. Witten, *Bioinformatics* **2004**, *20*, 2479–2481.

[81] J. Dearden, M. Cronin, K. Kaiser, *SAR and QSAR in Environmental Research* **2009**, *20*, 241–266.

1.7 Supplementary material



Supp. Fig. 1.1: Y-randomization test. Boxplot showing the Y-randomization test performance over 100 runs.

Table 1.6: Description of the subset of 120 selected genes.

KEGG Orthology	Gene	Definition	Functional category
K10233	aglF	alpha-glucoside transport system permease protein	ABC transporters
K05815	ugpE	sn-glycerol 3-phosphate transport system permease protein	ABC transporters
K11605	sitC	manganese/iron transport system permease protein	ABC transporters
K10234	aglG	alpha-glucoside transport system permease protein	ABC transporters
K09814	hrtA	hemin transport system ATP-binding protein	ABC transporters
K09687	ABC-2.AB.A	antibiotic transport system ATP-binding protein	ABC transporters
K02007	cbiM	cobalt/nickel transport system permease protein	ABC transporters
K02016	ABC.FEV.S	iron complex transport system substrate-binding protein	ABC transporters
K02049	ABC.SN.A, ssuB, tauB	sulfonate/nitrate/taurine transport system ATP-binding protein	ABC transporters
K11084	phnT	2-aminoethylphosphonate transport system ATP-binding protein	ABC transporters
K02073	ABC.MET.S, metQ	D-methionine transport system substrate-binding protein	ABC transporters
K10020	occQ, nocQ	octopine/nopaline transport system permease protein	ABC transporters
K05814	ugpA	sn-glycerol 3-phosphate transport system permease protein	ABC transporters
K09808	ABC.LPT.P, lolC, lolE	lipoprotein-releasing system permease protein	ABC transporters

K10228	smoF, mtlF	sorbitol/mannitol transport system permease protein	ABC transporters
K09813	hrtB	hemin transport system permease protein	ABC transporters
K10008	gluA	glutamate transport system ATP-binding protein	ABC transporters
K12372	dppF	dipeptide transport system ATP-binding protein	ABC transporters
K02051	ABC.SN.S, ssuA, tauA	sulfonate/nitrate/taurine transport system substrate-binding protein	ABC transporters
K02033	ABC.PE.P	peptide/nickel transport system permease protein	ABC transporters
K01997	livH	branched-chain amino acid transport system permease protein	ABC transporters
K10021	occP, nocP	octopine/nopaline transport system ATP-binding protein	ABC transporters
K10547	ABC.GGU.P, gguB	putative multiple sugar transport system permease protein	ABC transporters
K06074	ABC.VB12.A, btuD	vitamin B12 transport system ATP-binding protein	ABC transporters
K09692	tagG	teichoic acid transport system permease protein	ABC transporters
K10019	occM, nocM	octopine/nopaline transport system permease protein	ABC transporters
K10007	gluD	glutamate transport system permease protein	ABC transporters
K09970	aapQ, bztB	general L-amino acid transport system permease protein	ABC transporters
K10548	ABC.GGU.A, gguA	putative multiple sugar transport system ATP-binding protein	ABC transporters
K10227	smoE, mtlE	sorbitol/mannitol transport system substrate-binding protein	ABC transporters
K09996	artJ	arginine transport system substrate-binding protein	ABC transporters
K10230	smoK, mtlK	sorbitol/mannitol transport system ATP-binding protein	ABC transporters
K05776	modF	molybdate transport system ATP-binding protein	ABC transporters
K09971	aapM, bztC	general L-amino acid transport system permease protein	ABC transporters
K10440	rbsC	ribose transport system permease protein	ABC transporters
K11073	potF	putrescine transport system substrate-binding protein	ABC transporters
K10018	occT, nocT	octopine/nopaline transport system substrate-binding protein	ABC transporters
K11074	potI	putrescine transport system permease protein	ABC transporters
K10014	hisJ	histidine transport system substrate-binding protein	ABC transporters
K10006	gluC	glutamate transport system permease protein	ABC transporters
K11072	potA	spermidine/putrescine transport system ATP-binding protein	ABC transporters
K02407	fliD	flagellar hook-associated protein 2	Flagellar assembly, Motility
K02414	fliK	flagellar hook-length control protein FliK	Flagellar assembly, Motility
K02651	flp, pilA	pilus assembly protein Flp/PilA	Flagellar assembly, Motility
K00996	E2.7.8.6, rfbP	undecaprenyl-phosphate galactose phosphotransferase	LPS biosynthesis
K02851	rfe	undecaprenyl-phosphate alpha-N-acetylglucosaminyltransferase	LPS biosynthesis
K03273	gmhB	D-glycero-D-manno-heptose 1,7-bisphosphate phosphatase	LPS biosynthesis
K12996	rgpA	rhamnosyltransferase	LPS biosynthesis
K00912	lpxK	tetraacyldisaccharide 4'-kinase	LPS biosynthesis
K12988	wapR	alpha-1,3-rhamnosyltransferase	LPS biosynthesis
K13626	fliW	flagellar assembly factor FliW	Motility
K02383	flbB	flagellar protein FlbB	Motility
K07351	fimI	fimbrial protein	Motility
K02415	fliL	flagellar FliL protein	Motility
K07350	fimH	minor fimbrial subunit	Motility
K02662	pilM	type IV pilus assembly protein PilM	Secretion systems
K03110	ftsY	fused signal recognition particle receptor	Secretion systems
K02650	pilA	type IV pilus assembly protein PilA	Secretion systems
K11912	ppkA	serine/threonine-protein kinase PpkA	Secretion systems
K02453	gspD	general secretion pathway protein D	Secretion systems
K03116	tatA	sec-independent protein translocase protein TatA	Secretion systems
K03197	virB2	type IV secretion system protein VirB2	Secretion systems
K02281	cpaD	pilus assembly protein CpaD	Secretion Systems, Motility
K07345	fimA	major type 1 subunit fimbrin (pilin)	Secretion Systems, Motility
K02280	cpaC, rcpA	pilus assembly protein CpaC	Secretion Systems, Motility
K03219	yscC	type III secretion protein SctC	Secretion Systems, Motility
K11913	fha1	type VI secretion system protein	Secretion Systems, Motility
K02278	cpaA, tadV	prepilin peptidase CpaA	Secretion Systems, Motility
K02405	fliA	RNA polymerase sigma factor for flagellar operon FliA	Secretion systems, Motility
K01114	plcC	phospholipase C	Toxins
K11031	slo	thiol-activated cytolysin	Toxins
K08587	cloSI	clostripain	Toxins
K10954	zot	zona occludens toxin	Toxins
K01186	NEU1	sialidase-1	Toxins
K11041	eta	exfoliative toxin A/B	Toxins
K11018	tlh	thermolabile hemolysin	Toxins
K10953	rtxA	RTX toxin RtxA	Toxins
K11015	cdtC	cytolethal distending toxin subunit C	Toxins
K11032	hlyII	hemolysin II	Toxins
K11063	tcdAB	toxin A/B	Toxins

K07668	vicR	two-component system, OmpR family, response regulator VicR	Two-component
K10682	saeR	two-component system, OmpR family, response regulator SaeR	Two-component
K11712	dctR	two-component system, LuxR family, response regulator DctR	Two-component
K02480	K02480	two-component system, NarL family, sensor kinase	Two-component
K12973	pagP	palmitoyl transferase	Two-component
K13598	ntrY	two-component system, NtrC family, nitrogen regulation sensor histidine kinase NtrY	Two-component
K07645	qseC	two-component system, OmpR family, sensor histidine kinase QseC	Two-component
K07814	K07814	putative two-component system response regulator	Two-component
K07636	phoR	two-component system, OmpR family, phosphate regulon sensor histidine kinase PhoR	Two-component
K07768	senX3	two-component system, OmpR family, sensor histidine kinase SenX3	Two-component
K07652	vicK	two-component system, OmpR family, sensor histidine kinase VicK	Two-component
K11711	dctS	two-component system, LuxR family, sensor histidine kinase DctS	Two-component
K07704	lytS	two-component system, LytT family, sensor histidine kinase LytS	Two-component
K07653	mprB	two-component system, OmpR family, sensor histidine kinase MprB	Two-component
K13587	cckA	two-component system, cell cycle sensor histidine kinase and response regulator CckA	Two-component
K07685	narP	two-component system, NarL family, nitrate/nitrite response regulator NarP	Two-component
K08476	pgtA	two-component system, NtrC family, phosphoglycerate transport system response regulator PgtA	Two-component
K07716	pleC	two-component system, cell cycle sensor histidine kinase PleC	Two-component
K07705	lytT, lytR	two-component system, LytT family, response regulator LytT	Two-component
K07777	degS	two-component system, NarL family, sensor histidine kinase DegS	Two-component
K13599	ntrX	two-component system, NtrC family, nitrogen regulation response regulator NtrX	Two-component
K02475	K02475	two-component system, CitB family, response regulator	Two-component
K02667	pilR	two-component system, NtrC family, response regulator PilR	Two-component
K11527	K11527	two-component system, unclassified family, sensor histidine kinase and response regulator	Two-component
K13924	cheBR	two-component system, chemotaxis family, CheB/CheR fusion protein	Two-component
K07710	atoS	two-component system, NtrC family, sensor histidine kinase AtoS	Two-component
K13532	kinD	two-component system, sporulation sensor kinase D	Two-component
K07662	cpxR	two-component system, OmpR family, response regulator CpxR	Two-component
K02491	kinA	two-component system, sporulation sensor kinase A	Two-component
K02484	K02484	two-component system, OmpR family, sensor kinase	Two-component
K07637	phoQ	two-component system, OmpR family, sensor histidine kinase PhoQ	Two-component
K08475	pgtB	two-component system, NtrC family, phosphoglycerate transport system sensor histidine kinase PgtB	Two-component
K07638	envZ	two-component system, OmpR family, osmolarity sensor histidine kinase EnvZ	Two-component
K11357	divJ	two-component system, cell cycle sensor histidine kinase DivJ	Two-component
K07661	rstA	two-component system, OmpR family, response regulator RstA	Two-component
K07663	creB	two-component system, OmpR family, catabolic regulation response regulator CreB	Two-component
K07699	spo0A	two-component system, response regulator, stage 0 sporulation protein A	Two-component
K02476	K02476	two-component system, CitB family, sensor kinase	Two-component
K05875	tar	methyl-accepting chemotaxis protein II, aspartate sensor receptor	Two-component, Chemotaxis
K05874	tsr	methyl-accepting chemotaxis protein I, serine sensor receptor	Two-component, Chemotaxis

Table 1.7: List of misclassified organisms.

Organism	Taxon	Label	Classification	Percentaje
Acidobacterium capsulatum ATCC 51196	Acidobacteria	N	H	100
Tropheryma whipplei TW08/27	Actinobacteria	H	N	100
Acidimicrobium ferrooxidans DSM 10331	Actinobacteria	N	H	100
Mycobacterium leprae Br4923	Actinobacteria	H	N	50
Rhodobacter capsulatus SB 1003	Alphaproteobacteria	N	H	100
Erythrobacter litoralis HTCC2594	Alphaproteobacteria	N	H	100
Sphingobium japonicum UT26S	Alphaproteobacteria	N	H	75
Novosphingobium aromaticivorans DSM 12444	Alphaproteobacteria	N	H	85
Parvularcula bermudensis HTCC2503	Alphaproteobacteria	N	H	85
Rhodopseudomonas palustris BisB18	Alphaproteobacteria	N	H	95
Bordetella petrii DSM 12804	Betaproteobacteria	N	H	100
Haliangium ochraceum DSM 14365	Deltaproteobacteria	N	H	100
Campylobacter jejuni subsp. jejuni 81116	Epsilonproteobacteria	H	N	100
Campylobacter jejuni subsp. doylei 269.97	Epsilonproteobacteria	H	N	85
Sulfurimonas denitrificans DSM 1251	Epsilonproteobacteria	N	H	75
Lactococcus lactis subsp. cremoris SK11	Firmicutes	N	H	100
Staphylococcus aureus subsp. aureus MRSA252	Firmicutes	H	N	100
Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365	Firmicutes	N	H	100
Bacillus cereus ATCC 10987	Firmicutes	H	N	85
Staphylococcus aureus subsp. aureus Mu50	Firmicutes	H	N	90
Vibrio cholerae O1 biovar El Tor str. N16961	Firmicutes	H	N	50
Vibrio cholerae MJ-1236	Gammaproteobacteria	H	N	100
Yersinia pestis Angola	Gammaproteobacteria	H	N	100
Candidatus Blochmannia floridanus	Gammaproteobacteria	N	H	100
Escherichia coli S88	Gammaproteobacteria	N	H	100
Ferrimonas balearica DSM 9799	Gammaproteobacteria	N	H	100
Pseudomonas putida KT2440	Gammaproteobacteria	N	H	80
Pseudomonas putida F1	Gammaproteobacteria	N	H	55

Conclusión del Capítulo 1

El objetivo fundamental del trabajo presentado en el Capítulo 1 fue desarrollar una herramienta que permita predecir la patogenicidad de una bacteria exclusivamente a partir de información codificada en su genoma. Al momento de la publicación de este trabajo, no exisitían aproximaciones similares que permitieran realizar dicha tarea independientemente del grupo taxonómico al cual la bacteria de interés pertenenciera.

La definición de patogenicidad bacteriana no es algo que pueda establecerse exactamente, y el desarrollo de una infección bacteriana no depende solamente del microorganismo patógeno sino también del hospedador. Estos aspectos hacen que el modelo propuesto sea bastante simplista y no se ajuste totalmente a la complejidad que caracteriza a éste fenómeno. Por esta razón, solamente nos centramos en estudiar los patógenos humanos y descartamos aquellas bacterias que causan infecciones en otros organismos (por ejemplo, plantas). Sin embargo, contemplar una diversidad más amplia de hospederos (que reflejaría de forma más real el fenómeno biolǵico de la patogenicidad bacteriana) no hubiese sido posible ya que existe una carencia importante de meta-información de alta calidad en las bases de datos genómicas. En otras palabras, la enorme cantidad de genomas que son generados constantemente - en su mayoría - no poseen información asociada fácilmente extraíble y utilizable, por ejemplo fecha de aislamiento, región geográfica, hospedero, patogenicidad, etc.

Como perspectiva de este trabajo se pretende extender y actualizar el modelo de predicción para incorporar la capacidad de identificar patógenos para otros hospederos además del humano. Sin embargo, y de acuerdo a lo descrito anteriormente, es necesario establecer una forma automatizada y precisa de recuperar meta-información asociada a los más de 30,000 genomas bacterianos disponibles actualmente en bases de datos públicas. De esta necesidad se genera otra perspectiva en sí misma, que es la creación de una base de datos de meta-información generada a partir de la aplicación de herramientas de procesamiento de lenguaje natural sobre artículos científicos, que permitirá luego la utilización de diversos tipos de información para realizar predicciones y asociaciones genotipo-fenotipo.

2 Modeling the emergence of new pathogens from genomes

Citation:

Iraola G, Spangenberg L, Valenzuela S, Camargo A, Naya H* (2016) **Enhancing clinical microbiology by predicting the emergence of new pathogens from genomes**. *Unpublished*.

* Corresponding author

2.1 Abstract

Bacterial infections are the main threats for human health today, with more than 160 new emerging infections identified during the past 70 years. The consolidation of NGS technologies has increased the amount of completed genomes for both pathogenic and non-pathogenic bacteria, allowing the potential integration of clinical microbiology with bacterial genomics. The challenge lies in combining these two areas in order to improve bacterial infections control, by predicting the emergence of new pathogens. In this work we present the development of a mixed empirical-theoretical model that allows to predict the emergence of new pathogens just from the information coded in their genomes. For a certain pair of bacteria, the model simulates the horizontal transfer of virulence genes, which allows to predict the conversion from non-pathogen to pathogen. The model was applied to real bacterial communities characterized by NGS-metagenomics, calculating a risk index for the most probable new pathogens. Our results are the first steps towards a new chapter of integration between clinical microbiology and bioinformatics. In this work, we show that it is possible to combine these disciplines to study the emergence and evolution of new pathogens from a genomics perspective.

2.2 Introduction

Bacteria have lived on the Earth since 4.000 millions years; they probably were the first inhabitants of our planet [82]. Even since human beings appeared, they have coexisted and in many cases coevolved. Evidently the first bacteria were not human pathogens, so it is not trivial to mention that every existing human bacterial pathogen has emerged at least once during evolution. The present scenario is characterized by a small proportion of known bacterial diversity as common human pathogens (e.g: *Salmonella* Thypi, *Shigella dysenteriae* or *Streptococcus pneumoniae*), responsible for causing most worldwide threats to public health [83]. Additionally, the process of emergence and re-emergence of new pathogens, caused by the rapid evolution of bacterial genomes, is of major concern in clinical microbiology, since over

160 new emerging infections have been identified during the past 70 years [84]. The advent and consolidation of next generation sequencing technologies has dramatically increased the amount of completed bacterial genomes, for both known human pathogenic and non-pathogenic strains, and consequently expanded clinical microbiology towards new grounds. Nowadays the challenge lies in coupling next generation technologies and bioinformatics with current microbiological methods, to improve treatment and surveillance of bacteria that cause human diseases and potentially predict the emergence of new pathogens [85].

We have recently used a bioinformatics framework to assess the problem of predicting bacterial pathogenicity in humans [86]. In our previous work, we developed a machine learning classifier based on the information coded in bacterial genomes. More specifically, we took benefit from those completed genomes for which information of human pathogenicity or non-pathogenecity was available; for each genome we determined the presence/absence of more than 800 genes potentially related to pathogenicity. Based on a subset of 120 highly informative genes, we were able to train a Support Vector Machine (SVM) algorithm that classifies bacteria into human pathogens or non-pathogens, with a final accuracy of 95%. In addition to predict whether a certain bacteria may be pathogenic or non-pathogenic, we decided to extend the idea by exploring the gene repertories which classify bacteria into one of these classes. In other words, we were able to establish which combinations of these 120 genes determine pathogenic and non-pathogenic phenotypes. In this context, a complete new spectrum of questions arose and the keystone to their answer lies in the ability to determine which genes are involved in class shifting and how they are transferred between bacterial species.

When transformation from non-pathogenic to pathogenic is modeled by considering gene gain/loss events, the situation can be thought in the context of horizontal gene transfers (HGTs), which are well-known paradigms in bacterial evolution [87]. Horizontally acquired DNA might import new functions into bacteria and could confer a selective advantage to them, which in many cases might be associated with pathogenic phenotypes [88–91]. Considering that the pool of potential virulence genes present in a bacterial community is defined by species composition, a certain non-pathogenic bacteria will be able to

acquire genes that could contribute to its shift into pathogenic, depending on the genetic relatedness and relative abundance between species in the community. In the present work we propose a model that links empirical and theoretical aspects of bacterial evolution to simulate the emergence of human pathogens from non-pathogenic bacteria. First, for each non-pathogenic bacterium (hereinafter called acceptors) we have identified those genes in the set of 120 of the SVM whose acquirement implies the swap to human pathogen. Second, we have identified possible donors (bacteria that are natural carriers of these genes) and calculated a relative probability for each successful transfer event, based on several parameters such as physical distance between genes in the donor and compositional characteristics of both genomes. Hence, for each pair donor-acceptor we create a relative risk index of transformation into human pathogen, summarizing all possible transformation events between donor and acceptor. Finally, given the species composition of a sample, which can be easily assessed through NGS-metagenomics, for each non-pathogenic bacterium in the sample, we determine the total risk of it becoming a pathogen by weighting the relative risk index with the species abundance. Fig. 2.1 shows the main steps of our model; while in the present this model is just a basic and rather simplistic simulation exercise, with more available data it could start a completely new chapter of synergy between classic clinical microbiology and bioinformatics. In fact, extending the present model might open new avenues for the study of emerging pathogens evolution and make it possible to use genomic data in the improvement of surveillance, treatment programs and development of biotechnologies, with direct impact on global public health.

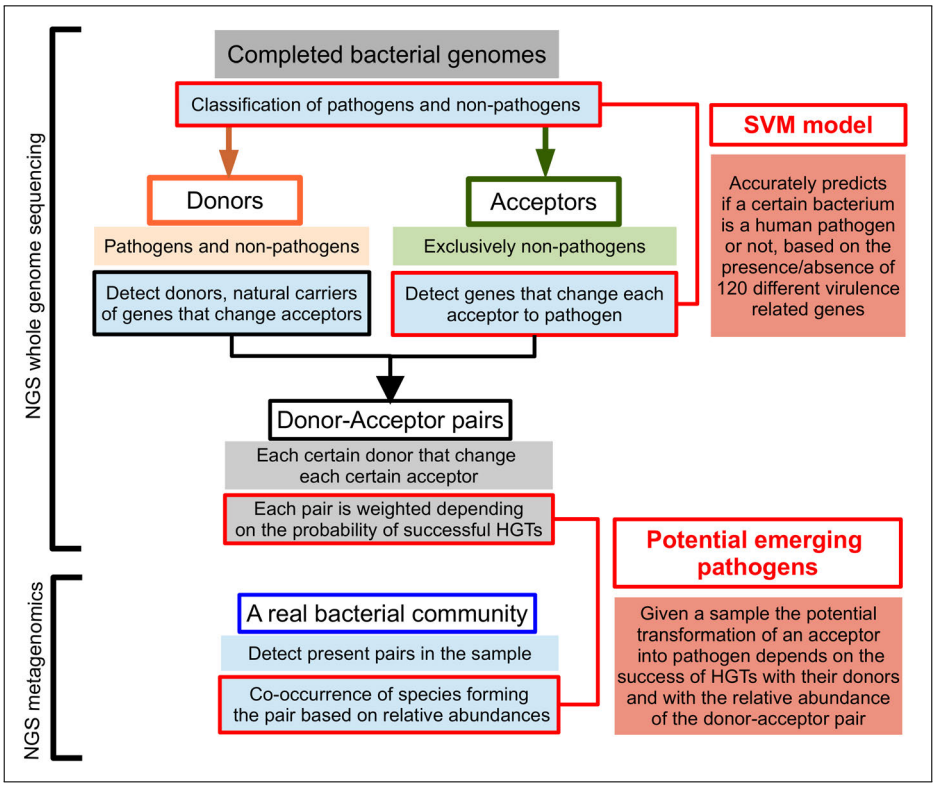


Figure 2.1: The global pipeline for PEPE. For completed bacterial genomes a SVM algorithm [86] is applied to classify them into human pathogens or non-pathogens. For each non-pathogen (acceptor) the set of genes which change it into pathogen is determined, then the group of bacteria which are natural carriers of these genes is established (donors). The probability for a certain acceptor to be turned into pathogen by a certain donor is based on the success of HGTs, which are simulated depending on compositional characteristics of genomes conforming the pair. For a real bacterial community, the co-occurrence of species conforming donor-acceptor pairs is taken into account to calculate the risk index for each acceptor, which can potentially be thought as a new emerging pathogen.

2.3 Results and Discussion

2.3.1 Phylogenetic representativeness. As the known bacterial biodiversity is clearly wider than the species represented by genomes (741) in our dataset, we assessed the representativeness of the included organisms comparing all bacterial taxonomic lineages from NCBI with lineages present in our dataset through Average Taxonomic Distinct-

ness (AvTD) and Variance Taxonomic Distinctness (VarTD) statistics (detailed in Methods).

The observed AvTD was 92.35 and ranged between 41.83 and 98.56. Based on re-sampling, the 95%-confidence interval for AvTD ranged between 91.39 and 92.10, which results in a significantly high observed AvTD. The observed VarTD was 208.00 and much lower than the 95%-confidence interval for VarTD, which ranged between 317.99 and 351.09. The significantly high AvTD and significantly low VarTD strongly suggest that the sample of bacterial lineages in the dataset is very representative of the whole known bacterial taxonomic diversity. In addition, the especially low VarTD suggests a uniform distribution or equal subdivision of sampled taxa among upper-level taxonomic categories. Therefore, the employed statistics suggest that our dataset represents an adequate sample of bacterial lineages for inference of evolutionary and phylogenetic patterns across all bacteria.

2.3.2 Modeling pathogens emergence. We exclusively define as potential acceptors those genomes belonging to non-pathogenic species that can be turned into pathogenic by horizontal acquisition of virulence genes present in other pathogenic or non-pathogenic species. We define as donors those genomes belonging to pathogenic or non-pathogenic species that are natural carriers of virulence genes that can be transferred to acceptors.

After running the model, 149 non-pathogenic genomes belonging to 124 different species (acceptors) were turned into pathogenic with probability greater than zero; on the other hand 241 non-pathogenic genomes were not changed to pathogenic in any case (this group is hereinafter called resilient), comprising 196 species. A total of 729 genomes belonging to 472 species (both pathogenic and non-pathogenic) performed as donors. This resulted in a matrix with dimensions $|729| \times |149|$ which defines all possible donor-acceptor pairs based on the calculated relative risk index for transformation from non-pathogen to pathogen. Further analysis of this index revealed remarkable biological features while comparing donors, acceptors and the resilient group.

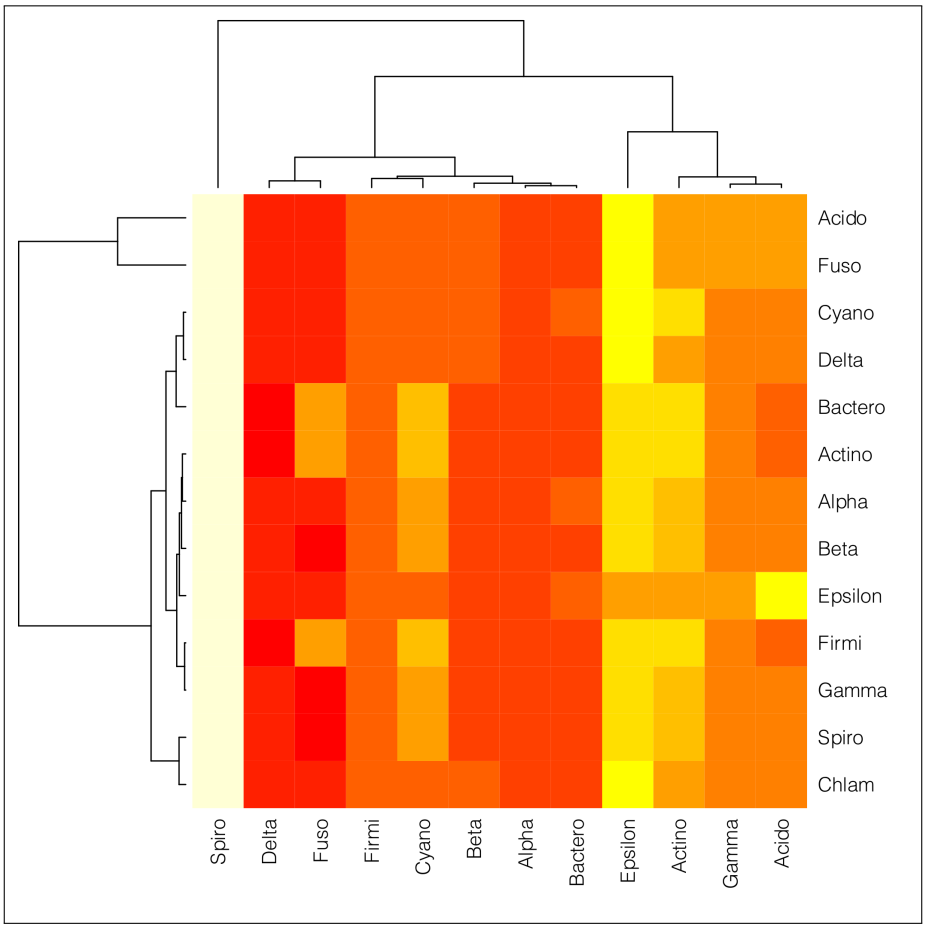


Figure 2.2: Heatmap of changed organisms per taxa. Represents the normalized number of bacteria changed by organisms of each taxonomic group. Donors are represented in the right-vertical while Acceptors in the bottom-horizontal axis. Color code determines the increasing number of changed organisms from Red (low) → White (high).

Donors were empirically classified as "strong" or "weak" based on their ability to transform non-pathogens into pathogens (Supp. Fig. 2.1), taking into account their accumulated risk index averaged over the number of acceptors that each donor was able to turn into a pathogen. For the purpose of our analysis we defined as "strong" the top 10% best performers and as “weak" the bottom 10% worst performers. The best donor resulted to be enterohemorragic *Escherichia coli* O157:H7 str. EDL933 and the remaining strong donors were

all well-known human pathogens, like *Salmonella enterica*, other *Escherichia coli* strains, *Vibrio cholerae*, *Yersinia pestis*, *Shigella flexneri*, *Streptococcus pneumoniae* and *Burkholderia mallei*; strong donors were limited to following taxa Gammaproteobacteria, Firmicutes and Betaproteobacteria. This suggests that species already adapted to human hosts are the most hazardous candidates to play a role in the emergence of new pathogens. On the contrary, weakest donors resulted to be mostly obligate parasites of vertebrates, like *Mycoplasma genitallium*, or insect endosymbionts like Candidatus *Blochmania floridanus*, also including a great number of marine and extremophiles bacteria. Weak donors include organisms from Actinobacteria, Alphaproteobacteria, Bacteroidetes/Chlorobi, Cyanobacteria, Deltaproteobacteria, Epsilonproteobacteria, Gammaproteobacteria, Firmicutes and Spirochaetes. In fact, these weak donors comprise a wider taxonomic range than strong donors and include a great proportion of non-pathogenic species (70%). Even though the model does not explicitly consider bacterial lifestyle as a parameter, results are extremely consistent on this subject, showing that obligate parasites, endosymbionts or any other organism living in extreme conditions (that impedes their physical contact with others) have an evident lower probability of performing as donors. In addition, these organisms are characterized by a small genome size and evident gene decay [92], which also contribute to their bad performance as donors.

Fig. 2.2 shows the normalized number of species, belonging to a certain taxon that are transformed into pathogenic by the other taxa. Acceptors are clearly divided in three groups: i) Spirochaetes are the most susceptible-to-transformation bacteria, meaning that there is a high probability that donors from all taxonomic groups may turn them into pathogens . ii) A similar scenario is observed in Epsilonproteobacteria and Actinobacteria, with a milder effect in Gammaproteobacteria and Acidobacteria; these taxa are represented by organisms with relatively high probability of being changed. iii) The last group is formed by hard-to-change organisms, in particular Deltaproteobacteria. A special case inside this group is Fusobacteria, which is particularly susceptible to Firmicutes, Bacteroidetes/Chlorobi and Actinobacteria.

Considering acceptors and the resilient group, Exact Fisher test was conducted to explore differences ($p < 0.05$) between these groups taking

their distributions among taxons into account. Significant differences were observed in Deltaproteobacteria, which are overrepresented in the resilient group ($p< 4 \times 10^{-4}$) and Actinobacteria, which are much more represented in acceptors ($p< 5 \times 10^{-5}$). This is coherent since it shows that Deltaproteobacteria, in addition of being the most hard-to-change taxon, is the group with greatest representation in the resilient group. On the contrary, Actinobacteria which is one of the most susceptible-to-transformation groups, is also characterized by a higher representation of acceptors when compared to the resilient group.

2.3.3 Genes implied in transformation. Horizontal gene transfer events imply the exchange of DNA fragments of variable sizes. In order to have an impact on the phenotype (e.g. pathogenicity) of the receptor bacteria this DNA fragment should carry at least one functional gene. However, the transfer of much more complex elements (like genomic islands carrying dozens of genes) is also possible [91]. Our model simulates three different situations: i) transfer of single genes, ii) pairs of genes and iii) three simultaneous genes.

In the first case, 60 out of 120 genes were able to change at least one genome into pathogenic. When analyzing functional categories (two-component and chemotaxis, ABC transporters, motility, secretion, toxins and LPS) of both groups of genes, only toxins were overrepresented ($p< 0.01$) in these genes that change compared to those that do not change. This is coherent because toxins have evolved to cause a direct pathogenic effect, while all other functional categories might be implied in pathogenicity but are also involved in basal cellular processes. Fig. 2.3 shows the number of acceptors changed by each single gene. The group of 60 genes that are able to change organisms to pathogenic is clearly divided in two subgroups: i) those which change almost all acceptors (14 genes) and ii) those which change a more restricted and reduced number of acceptors (46 genes). The first subgroup is dominated by Thiol-activated cytolysin (SLO), which is able to change all acceptors (toxin RTX, thermolabile hemolysin, toxin A/B and sialidase are also members of this subgroup), highlighting the exclusive role of toxins in pathogenesis and demonstrating that the incorporation of a single toxin-coding gene is enough to shift to the pathogenic phenotype. Additionally, strong differences arise when comparing genes

transferred by different donors (Supp. Fig. 2.2). On the one hand, weak donors transfer mainly ABC transporters genes (78%) and hardly transfer toxins, secretion systems and two-component genes. This bias in the genes-to-transfer repertory is coherent with the environmental constraints of weak donors, these organisms require ABC transporters for incorporation of essential metabolites which can not be produced due to gene decay, while toxins and other virulence-related genes are dispensable. On the other hand, strong donors transfer mainly what weak donors lack: toxins, secretion systems and two-component genes. These findings suggest that the donors' performance is strongly determined by the presence of genes belonging to these categories, pointing to their importance in bacterial pathogenicity.

Considering the transference of pairs of genes, 2092 out of 7240 (28%) possible combinations of 120 genes changed at least one genome into pathogenic. Further analysis of this subset of pairs revealed that 1773 (85%) were combinations of those 60 single genes that were able to change acceptors. However, we identified 6 genes (*gspD*, *envZ*, *creB*, *fliL*, *tatA* and *ftsY*) not belonging to the original subset of 60 single genes, which conform pairs that can actually change acceptors. A total of 319 pairs are formed by one of these 6 genes, and 6 pairs are formed by two of these 6 genes. This shows that even though some genes are not able to change acceptors exclusively by themselves, the combination with others unfolds new emergent properties that could turn the phenotype into pathogenic. A different scenario is observed with gene triplets. In this case the whole set of 120 genes is represented among those triplets that are able to change acceptors into pathogens. This reenforces the notion that the combination of more than 2 genes is auspicious for the emergence of new biological properties leading to pathogenic phenotype.

In summary, general trends are distinguishable among genes that are able to transform acceptors. As a general rule, the vast majority of single genes, pairs and triplets are able to change only a restricted number of acceptors into pathogens, while a small number of single genes, pairs and triplets change most acceptors.

2.3.4 Modeling real samples. The model was applied to analyze real bacterial communities assessed through NGS-metagenomics (re-

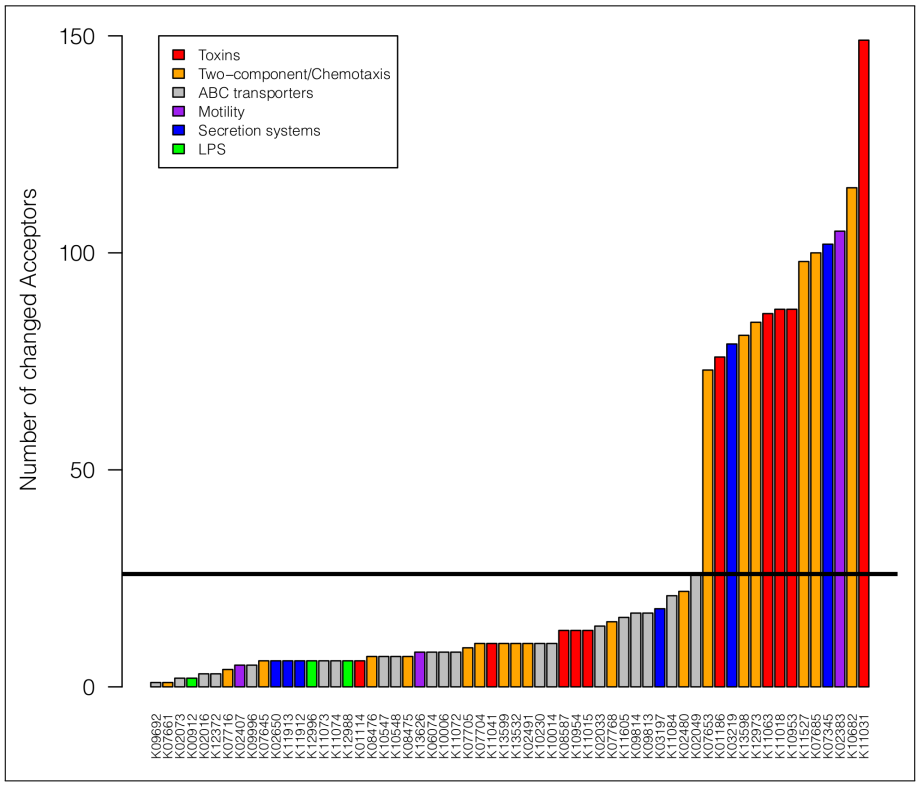


Figure 2.3: Barplot representing the number of Acceptors changed to pathogenic by each single gene transfer. Genes are named according to KEGG Orthology identifiers. Color code refers to the genes' functional categories: toxins (red), LPS (green), secretion systems (blue), two-component systems and chemotaxis (orange), motility (purple) and ABC transporters (grey). Horizontal black line defines 2 subgroups of genes according to the number of changed Acceptors (cut-off= 25).

trieved from MG-RAST database [93]), which provides high resolution of sample composition at a species level. Considering pairs of bacteria present in a certain metagenomic sample, which were also present in the model as pre-specified donor-acceptor pairs, the total risk index was calculated by weighting out the co-ocurrence of the pair in the sample. The whole sample was represented as a weighted and directed graph, where the nodes are bacteria and the edges (where the edge width is proportional to the risk index) represent the link between the donor and the acceptor in each pair. Those acceptors which are natural carriers of antibiotic-resistance plasmids were colored in blue, since

this feature could play a role in the success of an emerging pathogen.

At this point, the application of the model in real samples is limited by three main issues: i) the model is built based on a SVM classifier with an excellent (95%) but not optimal predictive performance, so initial misclassification of organisms as pathogens or non-pathogens could affect the ulterior result when simulating the emergence of new pathogens. ii) At this point of resolution we are not able to distinguish taxonomic levels lower than species. This implies that when many strains belonging to the same species are sequenced, we assume all of them are equally represented in the real sample. iii) The number of species present in the pre-computed dataset is restricted to those which are sequenced, precluding the analysis of the whole biodiversity that may exist in a sample. In spite of these disadvantages the results discussed below encourage the application of the model in real samples, while drawbacks will be gradually diminished as new completed bacterial genomes can be incorporated. This will allow to have a better representation of bacterial diversity and to improve the performance of classification and prediction models.

Human skin. Human skin is a proper environment for the development of both pathogenic and non-pathogenic microbiota [94]. This sample was mainly characterized by strong interactions between 4 species, with a higher number of weak background interactions (Fig. 2.4). We found that *Propionibacterium acnes* is the most powerful donor; this commensal bacteria is present in the skin of most healthy adults, but it is responsible for the pathogenic condition known as acne [95]. The second donor resulted to be *Streptococcus mitis*, also a commensal bacteria which commonly resides in the upper respiratory tract; however it is associated with endocarditis [96].

The highest risk index was observed between *P. acnes* and the acceptor *Cupriavidus metallidurans*. This bacterium has been extensively used in bioiengineering due to its tolerance to heavy metal stress [97], however it is not a common inhabitant of human skin and its presence is probably accidental. Another unusual inhabitant of human skin found in our sample is *Streptococcus thermophilus*, which presents

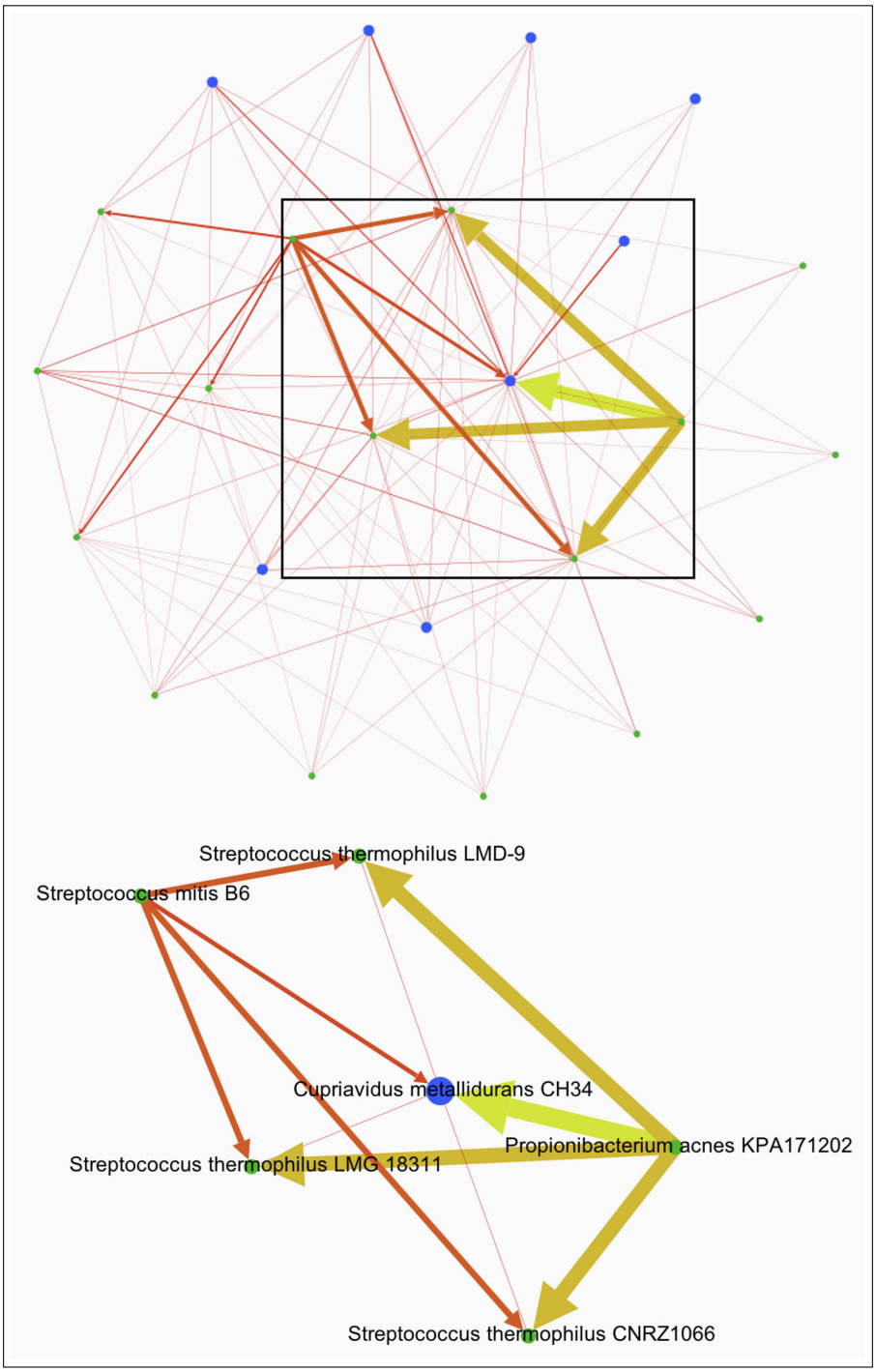


Figure 2.4: Output graph for skin sample. The complete graph is presented at the top, while at the bottom the figure is zoomed from the black rectangle, showing the most powerful interactions.

a high risk index with both *P. acnes* and *S. mitis*. These bacteria are usually found in farm environments and in the dairy industry, hence they are likely to be in contact with humans microbiota. Both situations point to the relevance of unusual species meddling in the community, considering that not only do *C. metallidurans* and *S. thermophilus* interact with *P. acnes*, but they also receive genes with considerable high probability from other 6 common inhabitants of the human skin (including *S. mitis*). Finally, on the one hand, *C. metallidurans* has been recently identified as the causative agent of septicemia, highlighting the underlying potential of this species as a possible emerging pathogen [98], which could be upgraded by the acquisition of virulence genes from the community. On the other hand, new insights in *S. thermophilus* genomics and evolution suggest a high HGT frequency between this species and other related pathogenic and non-pathogenic members of the *Streptococcus* genus [99].

Hospital air. Nosocomial-acquired infections are major threats for patients and still an unsolved problem in health-care. The analysis of a hospital air sample revealed a huge number of background and low probability interactions (involving 114 organisms); however, Fig. 2.5 shows a subset of organisms with a high probability of being involved in the emergence of new pathogens. In this situation, *Burkholderia multivorans* (a human pathogen associated to Cystic Fibrosis [100]) performs as the ringmaster, donating genes mainly to *Bacillus amyloliquefaciens*, *Corynebacterium efficiens* (carrying an antibiotic-resistance plasmid) and *Pseudomonas putida*. These three highly-susceptible bacteria are common inhabitants of soil and water, which facilites their transport and contact with human beings. This shows the importance of nosocomial settlements as hotspots of congregation between highly pathogenic and apparently harmless bacteria, allowing the interchange of their genomic features and the evolution of new pathogenic variants. In particular, *P. putida* (which is intimately related to pathogenic *P. aeruginosa* and *P. syringae*) has been reported to cause nosocomial infections [101], sustaining its importance as an emerging pathogen associated to hospital environments.

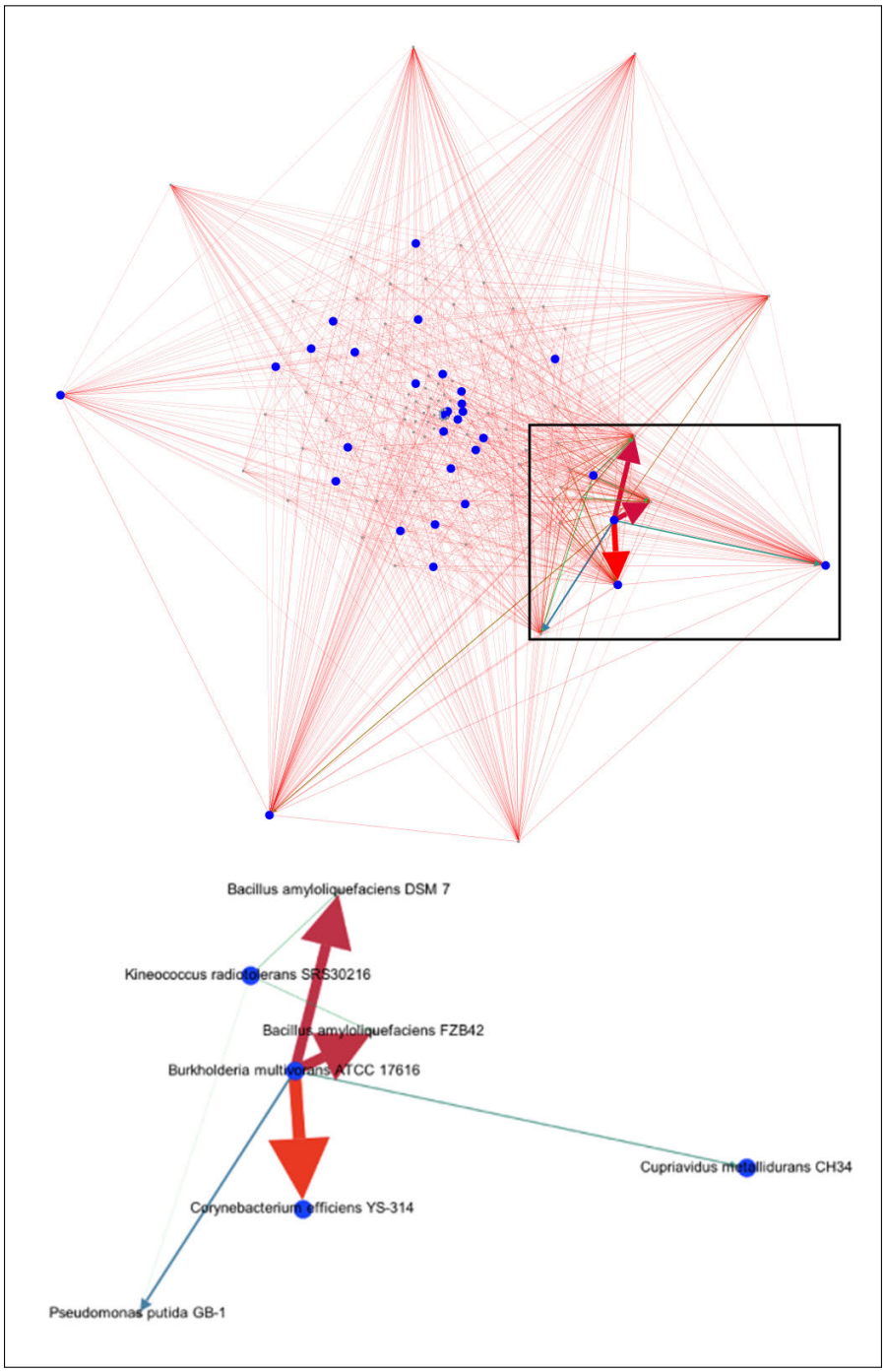


Figure 2.5: Output graph for hospital air sample. The complete graph is presented at the top, while at the bottom the figure is zoomed from the black rectangle, showing the most powerful interactions.

2.4 Conclusions

The consolidation of NGS technologies has opened new grounds of interaction between bacterial genomics and clinical microbiology. In our work, we integrate bioinformatics with theoretical and empirical aspects of bacterial and pathogenicity evolution to assess the prediction of new human pathogens. The landscape of emerging bacterial infections around the world is shaped by different socio-economics and ecological factors; it is currently possible to predict which combinations of these factors are more suitable for the emergence of new pathogens [84], but it is still difficult to predict the biological features of these new emerging bacteria. Here, we introduce the basic body of knowledge and the theoretical framework needed to address this issue and apply it in real life datasets.

Even though our model is simplistic (at this point) the results obtained are supported by documented biological evidence, proffering good perspectives for future work focused on incorporating more complex features. In this sense, improving simulations of HGTs by taking into account genomic environment (e.g, high recombination probability regions) and the presence of recombination and/or conjugation machineries in the donor-acceptor genomes will provide a more accurate prediction of horizontal gene transfer events. Additionally, the dynamics of bacterial communities are complex processes that not only depend on the species' relative abundances. The incorporation of population genetics and ecological models is undoubtedly necessary to enhance the prediction of emergent pathogens in different environmental conditions.

Our findings remark the critical need of strengthening the development of knowledge in the intersection between clinical microbiology and bioinformatics, as well as maintaining the rhythm of high accuracy whole-genome sequencing for pathogenic and non-pathogenic bacteria. Increasing the synergy between these grounds should result on direct improvements in the surveillance of emerging infectious diseases on a global scale.

2.5 Methods

2.5.1 Phylogenetic representativeness. We obtained a master list of 255271 records representing all bacterial taxonomic lineages in the NCBI Taxonomy Database. Each lineage in the list is classified into 8 taxonomic categories: phylum, class, subclass, order, suborder, family, genus, and species. From this list, we subsampled 741 records corresponding to the bacterial lineages with complete genomic sequences used in subsequent analyses. A Python-based script - PhyRep.phy - [102] was used to evaluate the phylogenetic representativeness of our sample in comparison with the known taxonomic diversity of Bacteria. This script implements the calculation of the Average and Variance Taxonomic Distinctness statistics (AvTD and VarTD respectively) [103, 104] of the sample corresponding to the average (or variance) path lengths among all pairs of tips (i.e., species in our case), based on the taxonomic tree. To test for significant AvTD and VarTD values, the script also randomly samples the master list to build a null distribution and 95%-confidence intervals for both statistics. We permuted the master list 1,000 times as suggested by the authors and used one-tailed tests based on the 95%-upper limit for AvTD and 95%-lower limit for VarTD. The analyses were repeated for a range of sampling sizes between 1,000 and 1,250 to explore the influence of sampling density via funnel-plots for AvTD and VarTD.

2.5.2 Defining genes that shift class to pathogen. Complete bacterial genome sequences were downloaded from the National Center for Biotechnology Information (NCBI). Over 1000 genomes were obtained; from these organisms we kept 741 which were labeled as human pathogens or non-pathogens. This set of bacteria comprehends 12 taxonomic groups (Acidobacteria, Fusobacteria, Gammaproteobacteria, *Alphaproteobacteria*, Epsilonproteobacteria, Betaproteobacteria, Deltaproteobacteria, Firmicutes, Spirochaetes, Actinobacteria, Bacteroidetes/Chlorobi, Chlamyidae/Verrucomicrobia). In this work we focused only on human pathogens; if a certain species was a multi-host pathogen which included humans; it was considered a human pathogen. On the contrary, if a certain species was a multi-host pathogen or a pathogen of other host different from human, it was ex-

cluded from the considered dataset. A previously described bacterial classification model [86] was used to determine gene gain/loss events that turn non-pathogenic bacteria into pathogenic. This model implements a Support Vector Machine algorithm based on the presence/absence of 120 virulence-related genes that comprise 6 functional categories (toxins, LPS biosynthesis, motility/flagellar assembly, secretion systems, two-component systems/chemotaxis and ABC transporters) to predict whether a certain genome will be human pathogenic or non-pathogenic. Moreover, once classified, the software allows to modify the presence/absence of each single gene considered and then reclassify the genome; this tool was used to record those situations in which non-pathogenic shifted to pathogenic considering the change in 1, 2 or 3 simultaneous genes. Once we determined which genes turn each non-pathogenic bacteria into pathogenic, we identified which bacteria (both pathogenic and non-pathogenic) are natural carriers of these genes and can serve as donors. Finally, we obtained all potential donor-acceptor pairs of bacteria.

2.5.3 Modeling HGT events. Our method is based on modeling HGT transfers of virulence-related genes between bacteria. Even though we identified all potential donor-acceptor pairs, this is not enough to ensure a successful HGT event. We further included several biophysical parameters to improve our calculations. In a first step, we took into account the DNA compositional characteristics of each donor-acceptor pair. A higher success of a particular HGT event itself is expected when it happens between organisms with similar codon usage, so we incorporated these parameters for each case. Eq. 1 describes an empirical estimator of matching codon usage (mcu) between 2 genomes, being $w1$ and $w2$ the vectors of relative usage for synonymous codons for donor and acceptor genome respectively.

$$mcu = \sqrt{1 - \frac{|\, w1 - w2\,|}{64}} \tag{1}$$

When simulating HGTs of 2 or 3 simultaneous genes, the physical distance between them was used to determine a success probability ($pdist$). We considered genomic islands (GIs) as the biggest known genetic mobile elements and used the frequency of GIs sizes along bac-

terial genomes to construct an empirical probability distribution, assuming that frequency of a certain GI size reflects the probability of an HGT event for this size. Supp. Fig. 2.3 shows the histogram of sizes for all well-characterized GIs retrieved from IslandViewer [105].

Furthermore, we considered the existing relation between the size of the transferred fragment and the size of the receptor genome in order to detect the maximum fragment that a certain genome could incorporate. We analyzed the relation between GIs and the carrier genome sizes and found no correlation between them, concluding that biological evidence supports the fact that a DNA fragment of any reasonable length can be inserted into any genome, independently of its size.

2.5.4 Transformation probability and theoretical risk index. To calculate the theoretical risk for a non-pathogenic bacteria to be transformed into a pathogenic one, we defined a metric based on two independent aspects: i) the probability that those genes capable of shifting a certain bacterium from non-pathogen to pathogen could be acquired via HGT, taking into account all parameters detailed above; ii) the probability that bacteria defining a donor-acceptor pair coexist in a natural environment. In this sense, we used MG-RAST database [93] to obtain relative abundance of bacterial species for samples of interest (assay repeatability was assessed by analyzing 5 samples of each experiment). Given two species belonging to a certain donor-acceptor pair, co-occurrence in the sample (dac) was calculated by multiplying dra and ara, corresponding to donor and acceptor relative abundances, respectively. Finally, the total risk index for a certain non-pathogenic bacteria was calculated by Eq. 2, depending on the number of transferred genes.

$$totalrisk = \begin{cases} mcu \times dac, & \text{if } genes = 1 \\ mcu \times dac \times pdist, & \text{if } genes > 1 \end{cases} \tag{2}$$

2.5.5 Model implementation in R and visualization of results. In order to define and implement the model in a stable environment, we developed an R package called PEPE (Pathogens Emergence Prediction Environment). For downstream analyses and interpretation of results, we chose graph representations implemented on Gephi [106].

Bacterial species are represented as graph nodes, while connections are directed and weighted edges that represent the value of the calculated risk index for any particular acceptor to be converted into pathogenic by a certain donor.

2.6 Acknowledgements

GI received financial support from Comisión Sectorial de Investigación Científica (CSIC), Uruguay. LS, SV and AC were supported by fellowships from the Agencia Nacional de Investigación e Innovación (ANII), Uruguay. HN was supported by grants from ANII (FCE.2.2011.1.7179) and Coordenação de Aperfeiçonamento de Pessoal de Nível Superior (CAPES), Programa Ciência Sem Fronteiras (PVE034.2012), Brazil. FOCEM Biomedicine.

2.7 References

[82] J. W. Schopf, *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 6735–6742.

[83] D. J. Ecker, R. Sampath, P. Willett, J. R. Wyatt, V. Samant, C. Massire, T. A. Hall, K. Hari, J. A. McNeil, C. Buchen-Osmond, B. Budowle, *BMC Microbiol.* **2005**, *5*, 19.

[84] K. E. Jones, N. G. Patel, M. A. Levy, A. Storeygard, D. Balk, J. L. Gittleman, P. Daszak, *Nature* **2008**, *451*, 990–993.

[85] X. Didelot, R. Bowden, D. J. Wilson, T. E. Peto, D. W. Crook, *Nat. Rev. Genet.* **2012**, *13*, 601–612.

[86] G. Iraola, M. Hernandez, L. Calleros, F. Paolicchi, S. Silveyra, A. Velilla, L. Carretto, E. Rodriguez, R. Perez, *J. Vet. Sci.* **2012**, *13*, 371–376.

[87] M. Touchon, P. Barbier, J. F. Bernardet, V. Loux, B. Vacherie, V. Barbe, E. P. Rocha, E. Duchaud, *Appl. Environ. Microbiol.* **2011**, *77*, 7656–7662.

[88] M. A. Tormo, E. Knecht, F. Gotz, I. Lasa, J. R. Penades, *Microbiology (Reading Engl.)* **2005**, *151*, 2465–2475.

[89] F. de la Cruz, J. Davies, *Trends Microbiol.* **2000**, *8*, 128–133.

[90] M. Juhas, D. W. Crook, D. W. Hood, *Cell. Microbiol.* **2008**, *10*, 2377–2386.

[91] M. Juhas, J. R. van der Meer, M. Gaillard, R. M. Harding, D. W. Hood, D. W. Crook, *FEMS Microbiol. Rev.* **2009**, *33*, 376–393.

[92] K. R. Sakharkar, P. K. Dhar, V. T. Chow, *Int. J. Syst. Evol. Microbiol.* **2004**, *54*, 1937–1941.

[93] F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, R. A. Edwards, *BMC Bioinformatics* **2008**, *9*, 386.

[94] E. A. Grice, H. H. Kong, G. Renaud, A. C. Young, G. G. Bouffard, R. W. Blakesley, T. G. Wolfsberg, M. L. Turner, J. A. Segre, *Genome Res.* **2008**, *18*, 1043–1050.

[95] H. Bruggemann, A. Henne, F. Hoster, H. Liesegang, A. Wiezer, A. Strittmatter, S. Hujer, P. Durre, G. Gottschalk, *Science* **2004**, *305*, 671–673.

[96] D. Denapaite, R. Bruckner, M. Nuhn, P. Reichmann, B. Henrich, P. Maurer, Y. Schahle, P. Selbmann, W. Zimmermann, R. Wambutt, R. Hakenbeck, *PLoS ONE* **2010**, *5*, e9426.

[97] P. J. Janssen, R. Van Houdt, H. Moors, P. Monsieurs, N. Morin, A. Michaux, M. A. Benotmane, N. Leys, T. Vallaeys, A. Lapidus, S. Monchy, C. Medigue, S. Taghavi, S. McCorkle, J. Dunn, D. van der Lelie, M. Mergeay, *PLoS ONE* **2010**, *5*, e10433.

[98] S. Langevin, J. Vincelette, S. Bekal, C. Gaudreau, *J. Clin. Microbiol.* **2011**, *49*, 744–745.

[99] C. Delorme, C. Bartholini, M. Luraschi, N. Pons, V. Loux, M. Almeida, E. Guedon, J. F. Gibrat, P. Renault, *J. Bacteriol.* **2011**, *193*, 5581–5582.

[100] J. J. Varga, L. Losada, A. M. Zelazny, L. Brinkac, D. Harkins, D. Radune, J. Hostetler, E. P. Sampaio, C. M. Ronning, W. C. Nierman, D. E. Greenberg, S. M. Holland, J. B. Goldberg, *J. Bacteriol.* **2012**, *194*, 6356–6357.

[101] G. Lombardi, F. Luzzaro, J. D. Docquier, M. L. Riccio, M. Perilli, A. Coli, G. Amicosante, G. M. Rossolini, A. Toniolo, *J. Clin. Microbiol.* **2002**, *40*, 4051–4055.

[102] F. Plazzi, R. R. Ferrucci, M. Passamonti, *BMC Bioinformatics* **2010**, *11*, 209.

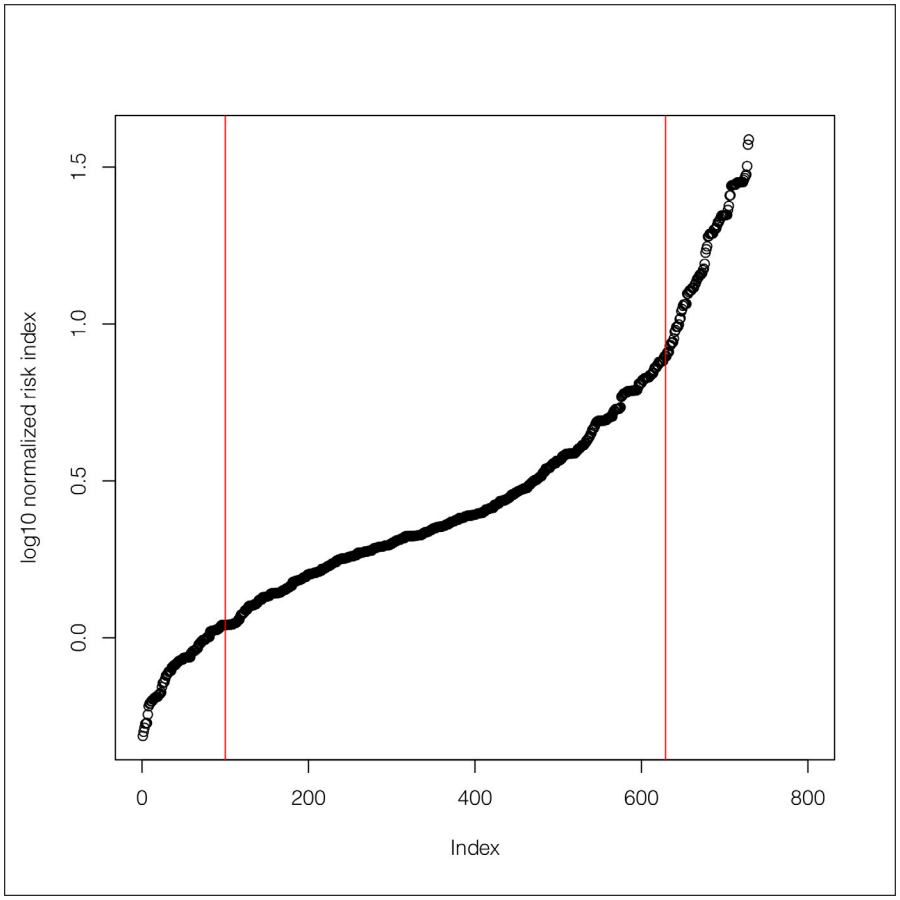
[103] K. R. Clarke, R. M. Warwick, *Journal of Applied Ecology* **1998**, *35*, 523–531.

[104] K. R. Clarke, R. M. Warwick, *Marine Ecology Progress Series* **2001**, *216*, 265–278.

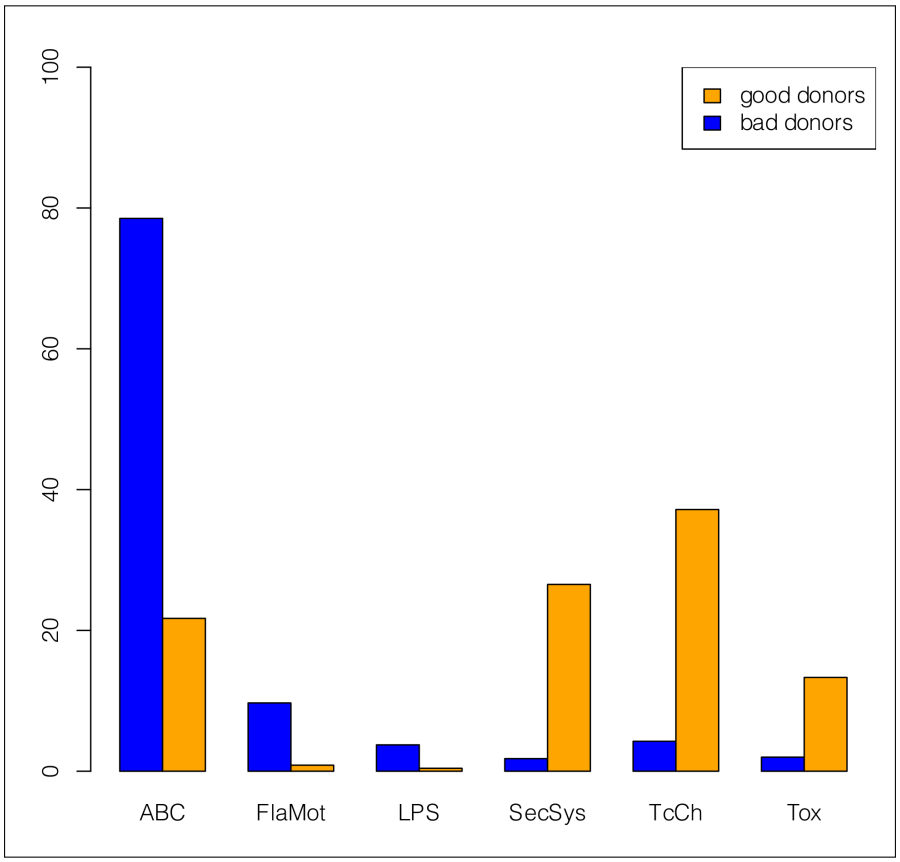
[105] M. G. Langille, F. S. Brinkman, *Bioinformatics* **2009**, *25*, 664–665.

[106] M. Bastian, S. Heymann, M. Jacomy, *International AAAI Conference on Weblogs and Social Media* **2009**.

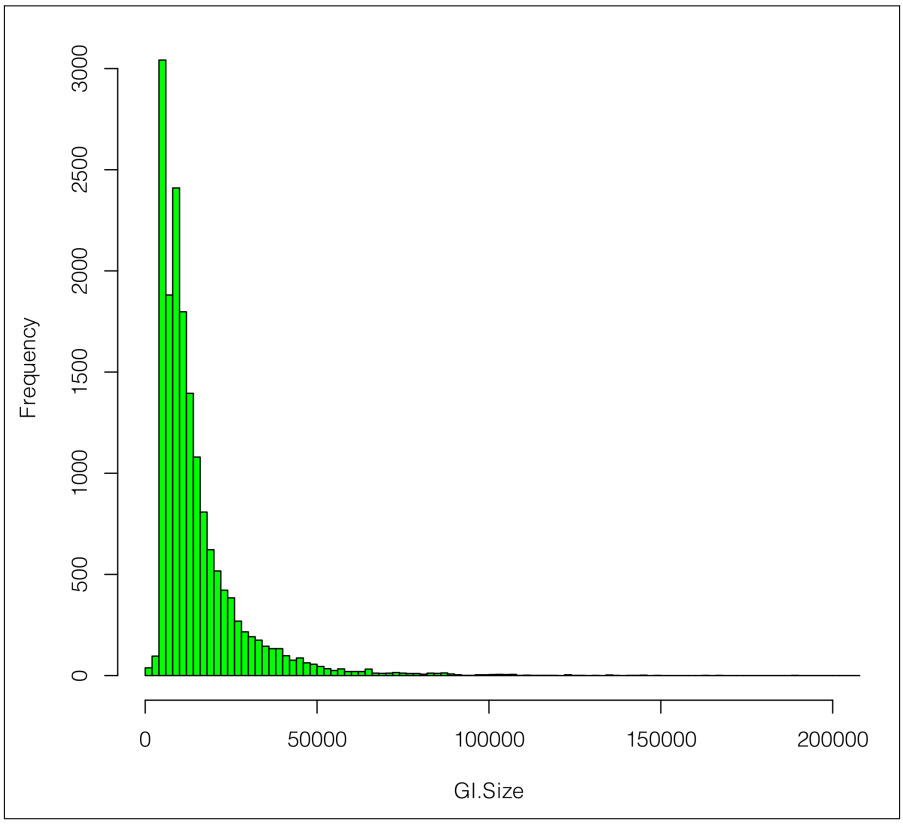
2.8 Supplementary material



Supp. Fig. 2.1: Normalized risk index. This figure shows the normalized risk index (log scale) calculated for each Donor in the model. Red vertical lines defines the 10% best Donors (rightside) and the 10% worst Donors (leftside).



Supp. Fig. 2.2: Transferred genes by donor. Barplot showing the percentage of transferred genes by weak and strong donors, separated by genes functional categories.



Supp. Fig. 2.3: Genomic islands size distribution. This figure shows the histogram for sizes of GIs retrieved from IslandViewer.

Conclusión del Capítulo 2

Este trabajo se da como consecuencia de los resultados descritos en el Capítulo 1, donde se identificaron un conjunto de genes altamente correlacionados con un fenotipo patogénico para humanos. Es así que conociendo estos genes es posible simular su pérdida o ganancia a través de eventos de transferencia horizontal y determinar si estos eventos son capaces de transformar una bacteria no patogénica en patogénica o viceversa.

Aplicamos esta aproximación en datos obtenidos por metagenómica para determinar la estructura de la comunidad bacteriana en diversos ambientes de interés sanitario. Mediante la aplicación de este modelo, es posible identificar en estas comunidades microorganismos donadores y aceptores de genes determinantes de patogenicidad, y simular los procesos de adquisición y pérdida que posibilitan la evolución y emergencia de nuevos patógenos.

La extensión de este modelo permitirá predecir en que condiciones, ambientes y ante la presencia de qué componentes dentro de una comunidad bacteriana es más riesgosa la emergencia de nuevos patógenos. Un aspecto clave a incorporar es la modelización de los eventos de transferencia horizontal en función de las características genómicas de los organismos donadores y aceptores. Por ejemplo, se puede asumir que un genoma con sistema CRISPR será más resistente a incorporar ADN foráneo que uno que no lo posee. De todas maneras, la aplicación del modelo tal cual esta descrito permitió identificar patógenos potencialmente emergentes cuyo riesgo real fue confirmado a través de búsquedas bibliográficas.

En los próximos años, y debido a la popularización de las tecnologías de secuenciación, será rutinario acceder y generar datos de secuenciación masiva para resolver casos clínicos. Para esto será clave el diseño de herramientas de identificación y predicción de patógenos a partir de los mismos.

3 Wedding higher taxonomic ranks with metabolic signatures coded in prokaryotic genomes

Citation:

Iraola G*, Naya H* (2016) **Wedding higher taxonomic ranks with metabolic signatures coded in prokaryotic genomes.** *Nature Microbiology*. Under review.

* Corresponding authors

3.1 Abstract

Taxonomy of prokaryotes has remained a controversial discipline due to the extreme plasticity of microorganisms, causing inconsistencies between phenotypic and genotypic classifications. The genomics era has enhanced taxonomy but also opened new debates about the best practices for incorporating genomic data into polyphasic taxonomy protocols, which are fairly biased towards the identification of bacterial species. Here we use an extensive dataset of Archaea and Bacteria to prove that metabolic signatures coded in their genomes are informative traits that allow to accurately classify organisms coherently to higher taxonomic ranks, and to associate functional features with the definition of taxa. Our results support the ecological coherence of higher taxonomic ranks and reconciles taxonomy with traditional chemotaxonomic traits inferred from genomes. KARL, a simple and free tool useful for assisting polyphasic taxonomy or to perform functional prospections is also presented (`https://github.com/giraola/KARL`).

3.2 Letter

IIn 1735, Carl von Linné released the Systema Naturae [107] setting a cornerstone in biology by establishing a formal system for the unambiguous nomenclature of living things, underpinning the sciences of taxonomy and systematics. However, these disciplines were originally conceived for eukaryotic organisms where classification was mainly based on morphological macroscopic traits, suffering from expectable inconsistencies when analogous principles were applied to prokaryotes. This caused that the classification prokaryotes remained a controversial field restricted to a small number of experts highly trained to develop and reproduce tuned-up chemotaxonomic and phenotypic tests. Afterwards, the pioneering work of Carl Woese in the mid-1980s caused a sudden change which shed light onto the evolutionary history of prokaryotes by introducing the 16S gene analysis [108]. Since then, many molecular characterization tools have been developed but the state-of-the-art strategy for assigning prokaryotes into novel or already described taxonomic units relies on polyphasic taxonomy, an approach

which tries to integrate all available chemotaxonomic and genotypic information to build a classification consensus [109].

The advent of high-throughput sequencing (HTS) has allowed the incorporation of genomic information into polyphasic taxonomy [110], but much effort has been invested to define classification rules for species in detriment of higher taxonomic ranks. Recently, two approaches that overcome this problem have been published: i) PhyloPhlAn constructs a high-resolution phylogenetic tree using a previously optimized set of 400 marker genes that accurately defines most taxa [111] and, ii) Microtaxi identifies taxon-specific genes and relies on a simple counting scheme to assign genomes to each taxon [112]. Both approaches were optimized with a subset of the available genomic diversity (around 2,000), do not provide automatic updating alternatives as new genomes become available and, fundamentally, do not allow to associate the taxonomic position with distinctive functional features of organisms.

In the present work we submit the hypothesis that higher taxonomic ranks (domain to genus) can be inferred from analyzing metabolic signatures of genomes. In turn, this hypothesis arises from the notion that higher ranks are ecologically coherent, meaning that most organisms within the same hierarchical level should display certain rationality in their lifestyles and ecological traits [113]. This is supported by the underlying signal of vertical evolution found in genes coding basal functions used as taxonomic markers [111, 114]. Indeed, signature metabolic genes that define taxon-specific ecological traits should be recognizable and used to define taxonomic ranks, similar to the approach implemented in Microtaxi [112], but also considering that the ecological coherence of high taxa could result from unique gene combinations, without any of them being taxon specific [113]. This requires the implementation of more powerful methods to capture informative patterns in highly-dimentional spaces.

To prove this ecological coherence of taxa at higher ranks and the usefulness of metabolic features as taxonomic markers, we used an extensive dataset of 33,236 archaeal and bacterial genomes representing 2 domains, 55 phyla, 67 classes, 163 orders, 328 families and 1,480 genera. For each genome the presence or absence of 1,328 different enzymes was assessed by parsing their annotations. Fig. 3.1a exemplifies

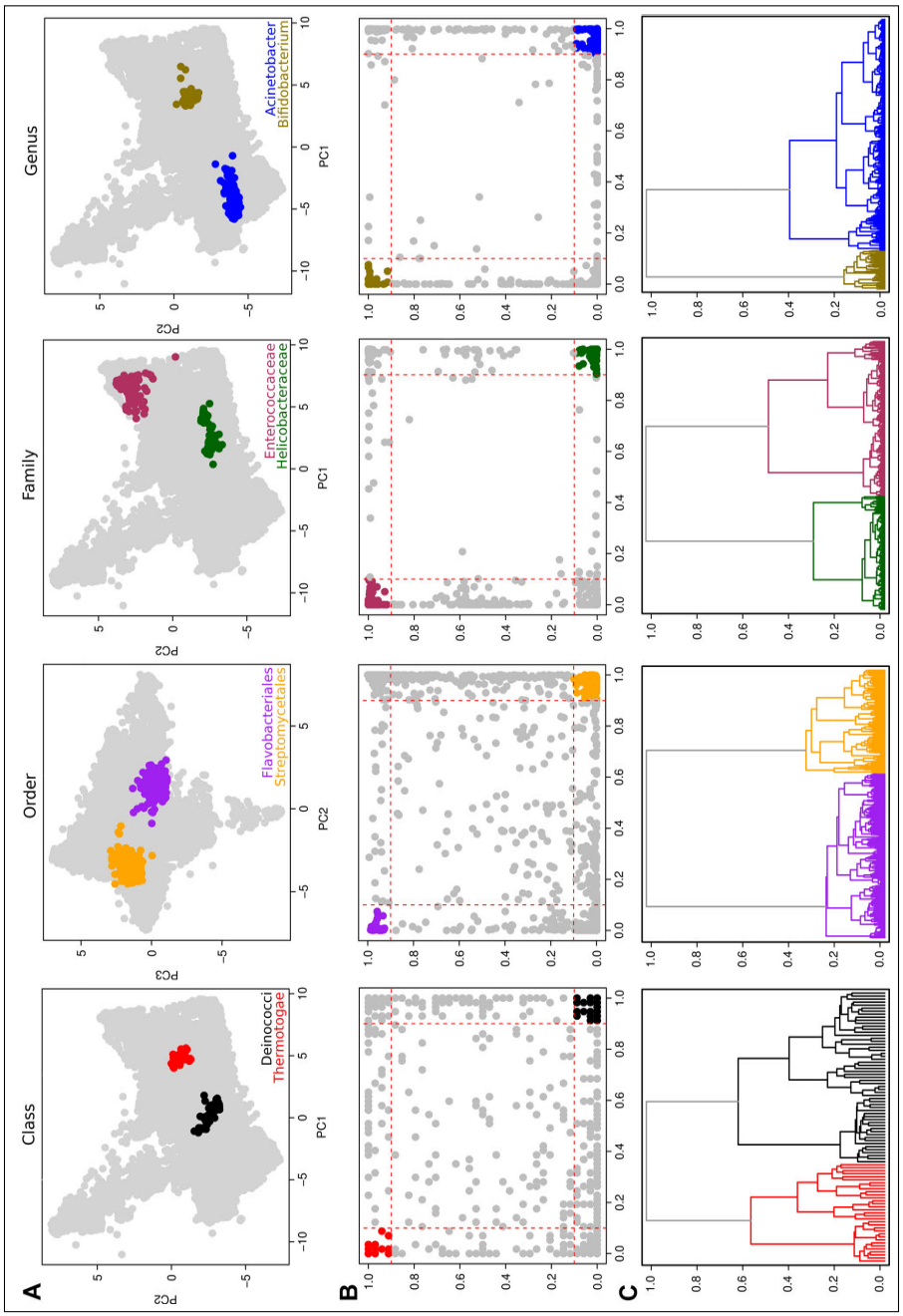


Figure 3.1: Informativeness of enzyme patterns. A) Principal Component Analysis using enzyme presence/absence vectors shows the discrimination of different pairs of taxa at every rank. B) For the same pairs of taxa, the frequency of each enzyme is plotted. Enzymes with frequency >0.9 in one taxon and <0.1 in the other and viceversa are highlighted. C) Cladogram based on pairwise Jaccard's distances.

how principal components analysis (PCA) resulting from enzyme patterns spatially discriminate taxa at different ranks. This separation is coherent with highly frequent or infrequent enzymes (Fig. 3.1b) and with Jaccard distance-based clustering (Fig. 3.1c). Then, we take the families Helicobacteraceae (intestinal gram-negatives associated with Crohn's disease [115]) and Enterococcaceae (intestinal gram-positives associated to probiotic effects[116]) to illustrate that enzyme patterns can cluster genomes according to their taxonomic position (Fig. 3.2a) based on the presence of distinctive combinations and subsets of marker enzymes (Fig. 3.2b). Beyond that identifying these single markers can provide useful information about taxon-specific molecular functions, we show that this information is also scalable to metabolic pathways allowing the isolation of those that significantly distinguish them (Fig. 3.2c). In Fig. 3.2d we show the full reference metabolism for starch and sucrose, which is one out of eight pathways that significantly distinguish both families each other (Fig. 3.2c), evidencing that the members of Enterococcaceae family present a vast distribution of enzymes while it is much more limited for the Helicobacteraceae genomes. This kind of functional prospection uncovers a strong link between metabolic potential and taxonomy, which indeed has been evidenced recently by modeling the variation of metabolomic data and community composition using metagenomic data [117].

As we showed that enzyme patterns are enough informative to discriminate taxa at different ranks and given the binary nature of this data, we built support vector machine (SVM) classification models using linear kernels by splitting the data in subsets corresponding to each taxa against the rest. All models were 10-fold corss-validated and performed very well independently of the taxonomic rank, reaching median precisions above 90%, false positive rates below 2% and false negative rates below 5% (Supp. Fig. 3.1). The very low rate of false positives is important since the practical cost of assigning a strain into a certain taxon is higher than keeping it as unknown. The explanation for non-optimal performance is the biased number of available genomes per taxa, since strong correlations ($R^2 = 0.76$, p-value $= 2 \times 10^{-16}$) were found between the number of genomes (at every rank) and classification performance measured as previously (Supp. Fig. 3.2). Additionally, when considering taxa with increasing

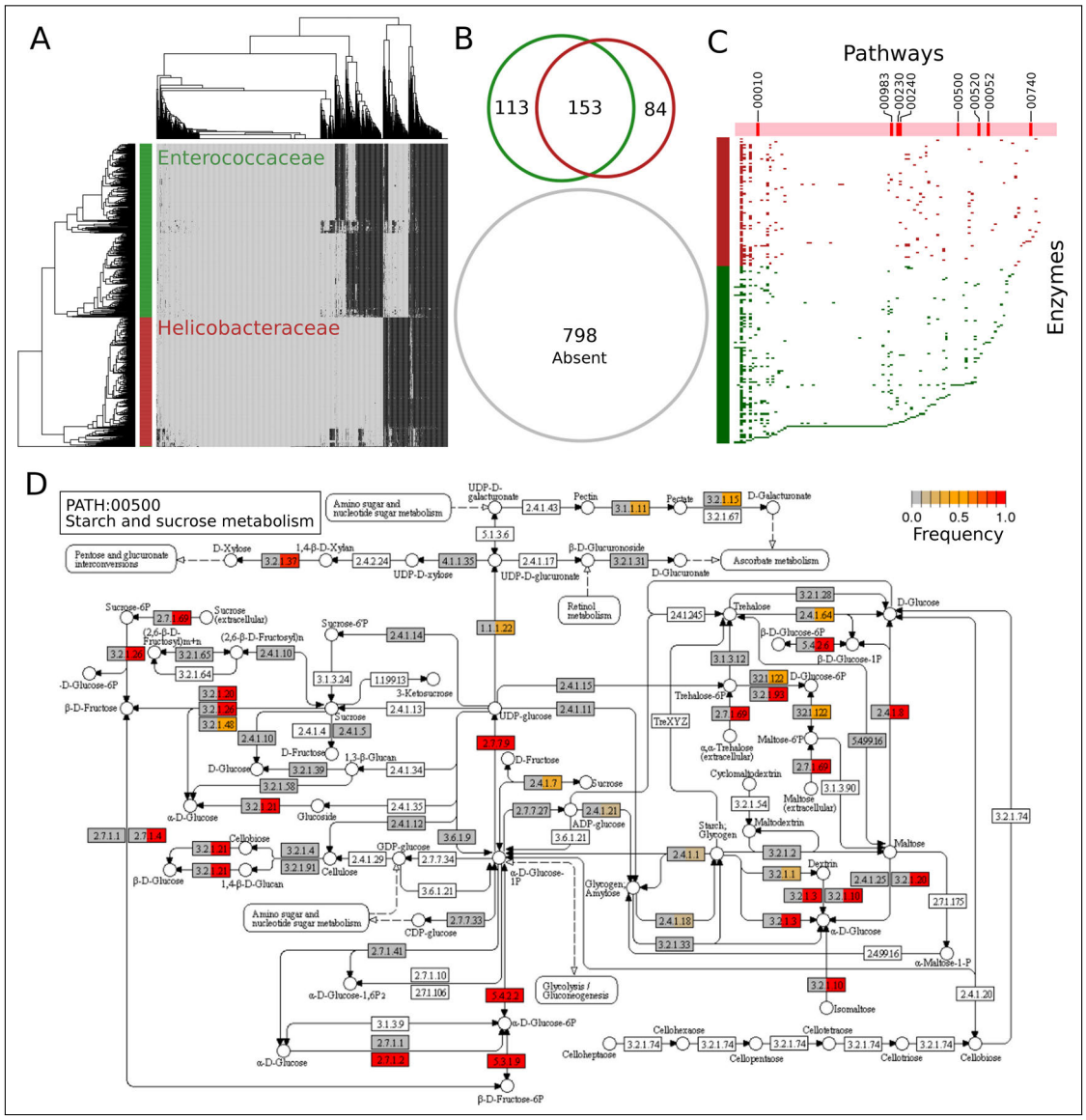


Figure 3.2: Example on families Helicobacteraceae and Enterococcacae. A) Heatmap showing the hierarchical clustering of genomes belonging to both families, exhibiting taxon-specific enzyme clusters. B) Venn diagram showing the distribution of enzymes between families. C) Distribution of enzymes inside metabolic pathways. Pathways that significantly differ (see Supp. Methods at `https://github.com/giraola/KARL/blob/master/userguide.pdf`) between families are labeled according to KEGG pathway identifiers. D) KEGG reference pathway for starch and sucrose metabolism. Each enzyme box is divided in two: left side for Helicobacteraceae and right side for Enterococcaceae. The frequency of each enzyme in each taxon is colored from grey (0) to red (1).

minimal number of genomes all performance parameters rapidly scale to optimal values. For example, when all taxa at family rank are included, the models showed a median precision of 94% and the interquartile range (IQR) between 72% and 100%, however when looking at models with at least 10 genomes the first quartile increased to 89% and the median to 95%, and for models with more than 50 genomes the first quartile scaled to 94% and the median to 97%. This tendency was observed for all ranks (Supp. Fig. 3.3) and demonstrated that classification errors are better explained due to biased data than to lack of information in enzyme patterns. Indeed, for some small taxa like classes Archaeoglobi (n = 7), Chlorobia (n = 12), Thermodesulfobacteria (n = 8) or Dictyoglomia (n = 2) classification precision was 100%. Interestingly, most of these taxa exhibit powerful ecological constraints reflected in very informative enzyme patterns that supports the ecological coherence of higher ranks and its association to genome-encoded signatures, overcoming their low representation in the whole dataset. Anyway, these observations reinforce the importance of sequencing genomes not only for the anthropocentric convenience but also for mere taxonomic interest, and also warrant the optimization of our approach as new genomes become available.

The robustness of predictions was further tested by applying the algorithm to an external set of 108 genomes obtained from very recent issues of Genome Announcements (`http://genomea.asm.org/`), hence not used in any step of model construction. At genus rank the average precision was 92% (Supp. Tab. 3.1), holding that obtained with cross-validated models. Interestingly, all misclassified genomes (n = 8) were predicted as unknown instead as any known genus erroneously, reinforcing the resilience of classification models against false positives. Additionally, the classification was totally consistent for genomes whose genus was truly unknown. For example, the bacilli bacterium VTI3104 was assigned within an unknown genus inside the Bacillaceae family, in accordance to its divergent phylogenetic position inside bacilli [118]. Alike, the Oscillatoriales cyanobacterium MTP1 was in fact classified within the Oscillatoriales order but was assigned to an unknown family and genus, in accordance to the low 16S identity (~90%) against its closest neighbors [119]. At higher ranks, classification was perfect for those genomes which were correctly assigned

at genus rank, since the algorithm takes benefit from the hierarchical structure of taxonomy and stops once the sample has been assigned to a certain genus by completing the lineage with precomputed information. Alike, when these genomes were tested for selected ranks above genus the results were totally coherent with the corresponding taxon. Seven out of 8 genomes that were incorrectly assigned at genus rank were then correctly classified at family rank. No misclassifications were observed at order, class, phylum and domain ranks.

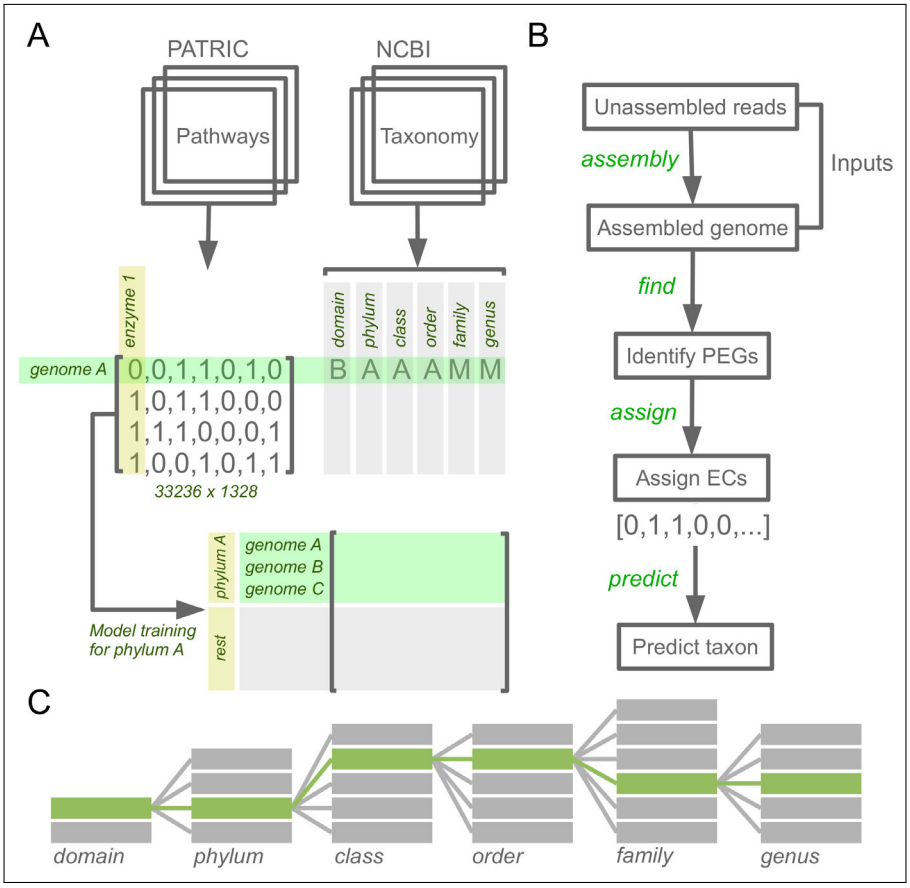


Figure 3.3: KARL pipeline. A) The workflow for building the dataset from genomes and annotations. B) Step-by-step analysis to classify a new genome. C) Schematic representation of included taxonomic ranks.

To ease the straightforward use of all models and datasets developed here, we built an R [120] package called KARL (`https://github.`

com/giraola/KARL) that allows the user to predict the membership of any newly sequenced genome to each high taxonomic rank by inputing just HTS reads, assembled genomes or annotation files. Additionally, the user can explore and improve each taxon-specific classification model by performing automatic feature selection procedures and compare between same rank taxa. A handful of functions allows the identification, comparison and visualization of meaningful metabolic signatures among taxa, including the automated production of all graphs and illustrations herein presented. Finally, it can connect the PATRIC database [121] to automatically update models and datasets based on newly released genomes. The full KARL pipeline is shown in Fig. 3.3. A detailed user guide explains theoretical and practical aspects from installation to step-by-step implementation of all features herein described (Supp. Methods at https://github.com/giraola/KARL/blob/master/userguide.pdf). KARL is intended to be a useful tool for assisting polyphasic taxonomy and to perform functional prospections and comparisons of prokaryotic genomes.

3.3 Methods

3.3.1 Genomic data. Annotation files were accessed from the PATRIC database [121] for a total of 33,236 available prokaryotic genomes. In-house R [120] scripts were designed to extract enzymes and pathways data from each file and build a presence/absence matrix; each column in the matrix represents one out of 1,328 different enzymes detected in at least one genome and identified with its unique EC number. The taxize R package [122] was used to retrieve the domain, phylum, class, order, family and genus for each genome from the NCBI Taxonomy database.

3.3.2 Classification models. RWeka [123] was used to train Support Vector Machine (SVM) models by splitting the data in two categories: one containing the considered taxon and other with the rest. For all cases, given the binary nature of data, a linear kernel function was preferred [124]. Each model can be improved by performing a feature selection scheme that involved three steps: i) attributes (enzymes) with a frequency lower than 0.1 or greater than 0.9 in both

categories were removed, ii) over the remaining set highly correlated attributes (>0.9) were removed just keeping those with lower average correlation values and, iii) Information Gain ratios were calculated for each attribute and a Davies test for slope change was applied to identify the Gain Ratio cut-off for keeping those most informative enzymes (Supp. Methods). Evaluation was initially performed by implementing a 3-repeated 10-fold cross-validation scheme to each model and then misclassified genomes were evaluated using a metric based on the integration of Hamming distances in the PCA space (Supp. Methods). Further testing was performed with an external dataset (Supp. Tab. 3.1).

3.3.3 Taxonomy prediction. The first step for predicting the membership of a new genome into any taxonomic unit implies to determine its presence/absence pattern for the 1,328 enzymes tested. Assessing this depends on the input: for unassembled sequencing reads the SPAdes genome assembler [125] generates a *de novo* assembly, else, if the input is a draft or finished genome the pipeline starts using Prodigal [126] and/or Glimmer [127] to predict protein-encoding genes. Finally, BLASTp [79] identifies enzymes presence/absence by searching against individual databases built from FIGfams [128] and KEGG Orthology [129]. A certain enzyme is considered present if there is any BLASTp hit with identity >70%, query coverage >90% and e-value <0.001 (these values were selected based on a grid search analysis that evaluated the classification performance using combinations of identity from 25% to 95% and query coverage from 50% to 95%) (Supp. Fig. 3.4). The presence/absence vector is then inputed to each taxon-specific SVM model.

3.3.4 KARL package. The whole methods were implemented as an R package called KARL freely available at GitHub repository (`https://github.com/giraola/KARL`). This package allows the standalone implementation of the whole applications described here in three operational modules: Explorer, Predictor and Updater. Explorer implements a handful of functions for automatic comparison of taxa, allowing to identify metabolic signatures at enzyme and pathway levels. Predictor allows to predict taxonomy from sequencing reads, assembled

or annotated genomes, evaluate predictions and optimize classification models. Updater allows to automatically update datasets and models with new available sequenced genomes in public databases. An in-depth description of practical and theoretical aspects are provided in the full user manual (Supp. Methods).

3.4 Acknowledgments

3.5 References

[79] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **1990**, *215*, 403–410.

[107] C. v. Linneaus, *Systema naturae. Vol. 1*, Holmiae :Impensis Direct. Laurentii Salvii, **1758**, p. 881.

[108] C. R. Woese, O. Kandler, M. L. Wheelis, *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 4576–4579.

[109] P. Vandamme, B. Pot, M. Gillis, P. de Vos, K. Kersters, J. Swings, *Microbiol. Rev.* **1996**, *60*, 407–438.

[110] M. Kim, H. S. Oh, S. C. Park, J. Chun, *Int. J. Syst. Evol. Microbiol.* **2014**, *64*, 346–351.

[111] N. Segata, D. Bornigen, X. C. Morgan, C. Huttenhower, *Nat Commun* **2013**, *4*, 2304.

[112] A. Gupta, V. K. Sharma, *BMC Genomics* **2015**, *16*, 396.

[113] L. Philippot, S. G. Andersson, T. J. Battin, J. I. Prosser, J. P. Schimel, W. B. Whitman, S. Hallin, *Nat. Rev. Microbiol.* **2010**, *8*, 523–529.

[114] B. Snel, P. Bork, M. A. Huynen, *Nat. Genet.* **1999**, *21*, 108–110.

[115] N. O. Kaakoush, J. Holmes, S. Octavia, S. M. Man, L. Zhang, N. Castano-Rodriguez, A. S. Day, S. T. Leach, D. A. Lemberg, S. Dutt, M. Stormon, E. V. O'Loughlin, A. Magoffin, H. Mitchell, *Helicobacter* **2010**, *15*, 549–557.

[116] S. Lory, *The Family Enterococcaceae*, *Vol. 1*, Springer, **2014**, pp. 75–77.

[117] C. Noecker, A. Eng, S. Srinivasan, C. M. Theriot, V. B. Young, J. K. Jansson, D. N. Fredricks, E. Borenstein, *mSystems* **2016**, *1*, (Ed.: L. M. Sanchez), DOI 10.1128/mSystems.00013-15.

[118] G. Tetz, V. Tetz, *Genome Announc* **2015**, *3*.

[119] P. C. Hallenbeck, M. Grogger, M. Mraz, D. Veverka, *Genome Announc* **2016**, *4*.

[120] R Development Core Team, R: A Language and Environment for Statistical Computing, ISBN 3-900051-07-0, R Foundation for Statistical Computing, Vienna, Austria, **2008**.

[121] A. R. Wattam, D. Abraham, O. Dalay, T. L. Disz, T. Driscoll, J. L. Gabbard, J. J. Gillespie, R. Gough, D. Hix, R. Kenyon, D. Machi, C. Mao, E. K. Nordberg, R. Olson, R. Overbeek, G. D. Pusch, M. Shukla, J. Schulman, R. L. Stevens, D. E. Sullivan, V. Vonstein, A. Warren, R. Will, M. J. Wilson, H. S. Yoo, C. Zhang, Y. Zhang, B. W. Sobral, *Nucleic Acids Res.* **2014**, *42*, D581–591.

[122] S. A. Chamberlain, E. Szocs, *F1000Res* **2013**, *2*, 191.

[123] K Hornik, A Zeileis, T Hothorn, C Buchta, *R package version 0.3-2* **2007**.

[124] G. Iraola, G. Vazquez, L. Spangenberg, H. Naya, *PLoS ONE* **2012**, *7*, e42144.

[125] A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev, P. A. Pevzner, *J. Comput. Biol.* **2012**, *19*, 455–477.

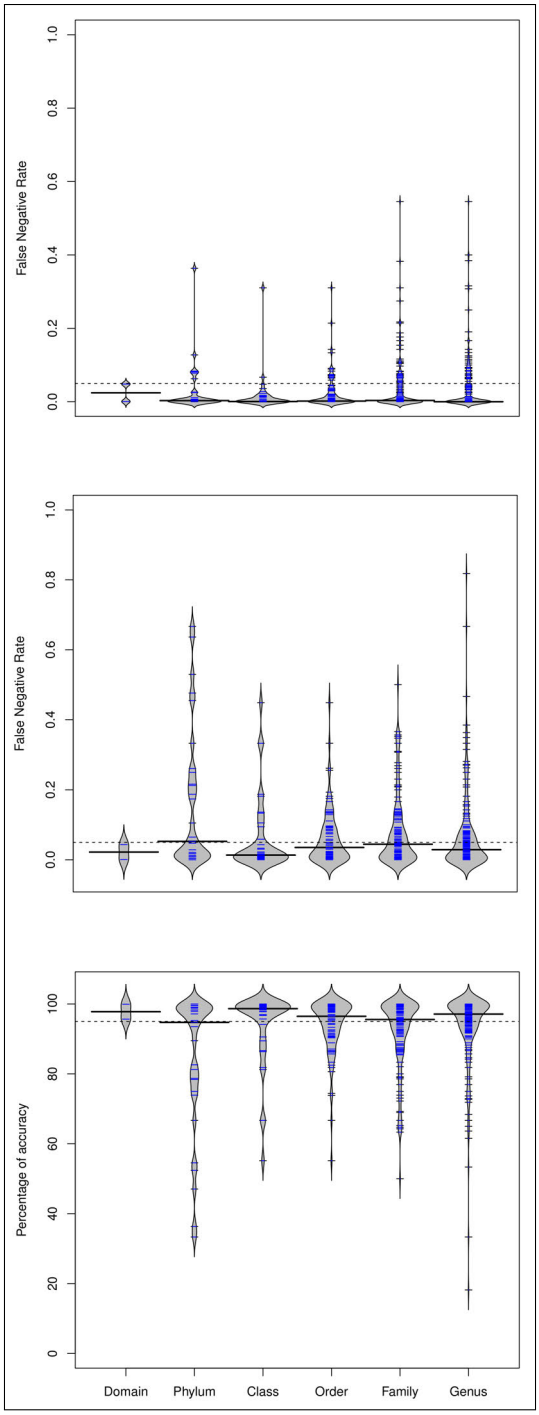
[126] D. Hyatt, G. L. Chen, P. F. Locascio, M. L. Land, F. W. Larimer, L. J. Hauser, *BMC Bioinformatics* **2010**, *11*, 119.

[127] A. L. Delcher, D. Harmon, S. Kasif, O. White, S. L. Salzberg, *Nucleic Acids Res.* **1999**, *27*, 4636–4641.

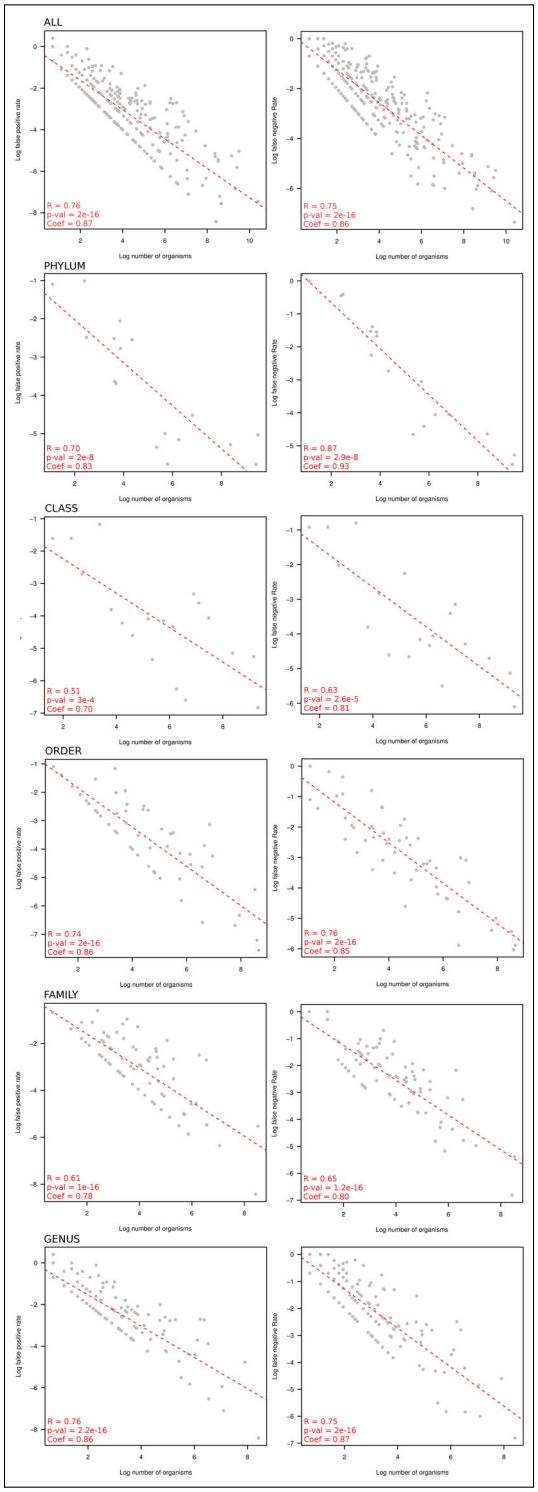
[128] F. Meyer, R. Overbeek, A. Rodriguez, *Nucleic Acids Res.* **2009**, *37*, 6643–6654.

[129] M. Tanabe, M. Kanehisa, *Curr Protoc Bioinformatics* **2012**, *Chapter 1*, Unit1.12.

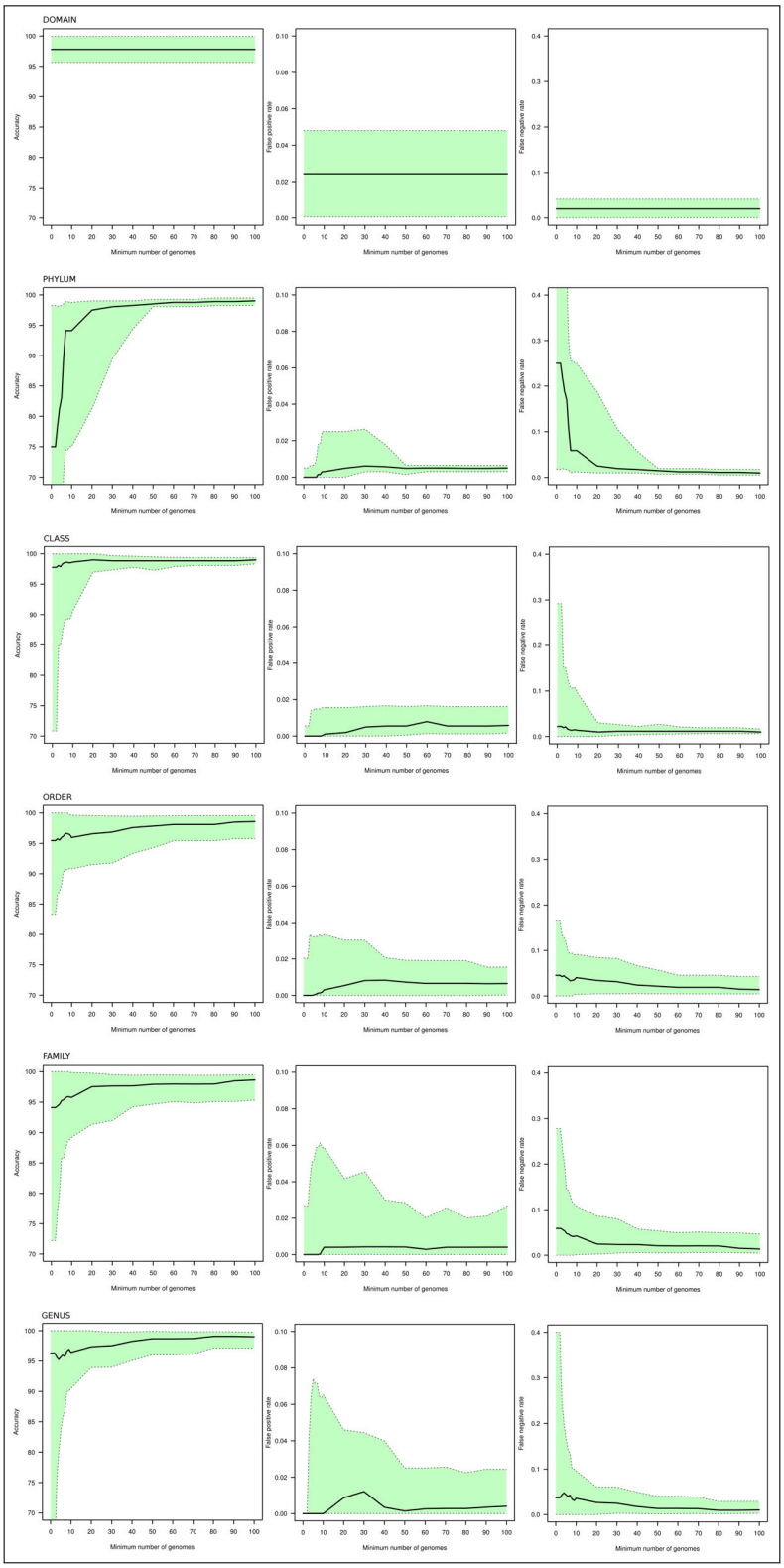
3.6 Supplemental material



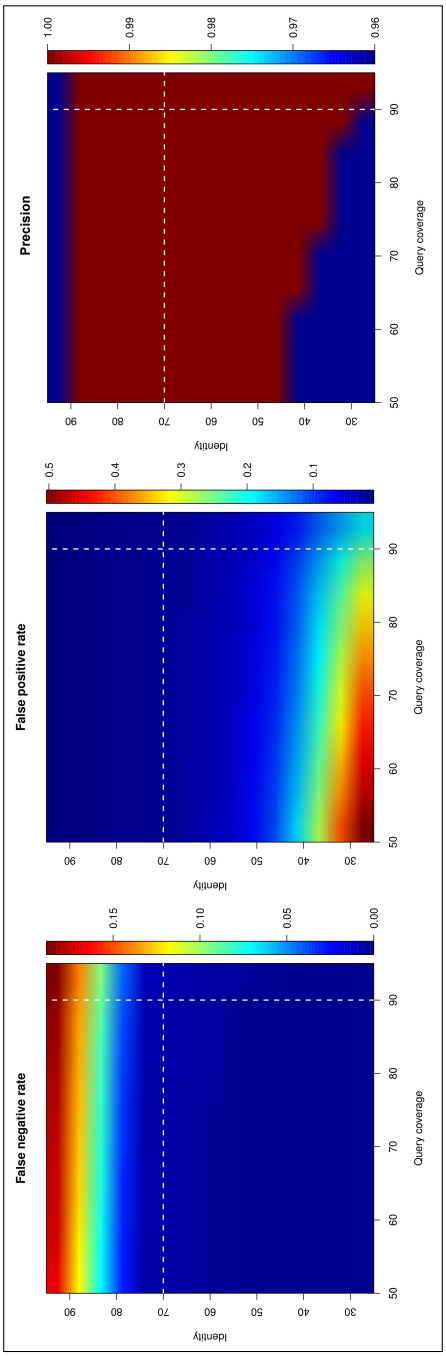
Supp. Fig. 3.1: Performance of classification models at every taxonomic rank. A) Distribution of false positive rates. B) Distribution of false negative rates. C) Distribution of precisions.



Supp. Fig. 3.2: Correlations between error rates and taxon sizes. Linear regressions showing correlations between false negative and false positive rates and the number of organisms per taxon at each taxonomic rank.



Supp. Fig. 3.3: Classification performance at different taxon sizes. Classification performance (precision, false positive and false negative rates) measured in taxa with increasing number of genomes. The black line is the median and the green dispersion is according to the upper and lower interquartile range boundaries.



Supp. Fig. 3.4: Grid search analysis. All possible combinations of BLAST hit identity (25% to 95%) and query coverage (50% to 95%) were assayed and for each result the false negative rate, false positive rate and precision were plotted. The dashed white lines indicate identity = 70% and query coverage = 90%, these values were set as default since they minimize error rates and maximizes precision.

Supp. Tab. 3.1: External set of test genomes. For column "prediction": Green: correct known, blue: correct unknown, red: incorrect.

genome	domain	phylum	class	order	family	genus	prediction
alphaproteobacterium_LFTY0	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Unknown	
candidatus_nasuia_deltocephalinicola_PUNC_CP013211	Bacteria	Proteobacteria	Gammaproteobacteria	Unknown	Unknown	Unknown	
candidatus_sulcia_muelleri_PUNC_CP013212	Bacteria	Bacteroidetes	Flavobacteriia	Flavobacteriales	NA	Candidatus Sulcia	
deinococcus_grandis_ATCC43672_BCMS0	Bacteria	Deinococcus-Thermus	Deinococci	Deinococcales	Deinococcaceae	Deinococcus	
dvosia_sp_H5989_CP011300	Bacteria	Proteobacteria	Unknown	Unknown	Unknown	Unknown	
micromonospora_RV43_LEKG0	Bacteria	Actinobacteria	Actinobacteria	Micromonosporales	Micromonosporaceae	Micromonospora	
nitrosomonas_communis_Nm2_CP011451	Bacteria	Proteobacteria	Betaproteobacteria	Nitrosomonadales	Nitrosomonadaceae	Unknown	
nocardia_seriolae_U1_BBYQ0	Bacteria	Actinobacteria	Actinobacteria	Corynebacteriales	Nocardiaceae	Nocardia	
nocardiopsis_RV163_LEKI01	Bacteria	Actinobacteria	Actinobacteria	Streptosporangiales	Nocardiopsaceae	Nocardiopsis	
oscillatoriales_cyanobacterium_MTP1_LNAA0	Bacteria	Cyanobacteria	NA	Oscillatoriales	Unknown	Unknown	
acetobacter_tropicalis_NBRC16470_BBMU0	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Acetobacteraceae	Acetobacter	
achromobacter_xylosoxidans_CF304_LFHA01	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Alcaligenaceae	Achromobacter	
acinetobacter_baumannii_Ab5_LANH000000000	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Acinetobacter	
actinobacillus_pleuropneumoniae_1022_JSVF0	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Actinobacillus	
aeromonas_caviae_L12_JWJP01	Bacteria	Proteobacteria	Gammaproteobacteria	Aeromonadales	Aeromonadaceae	Aeromonas	
alphaproteobacterium_Q1_BAYV01	Bacteria	Proteobacteria	Alphaproteobacteria	Unknown	Unknown	Unknown	
amycolatopsis_orientalis_CPCC200066_JXRD01	Bacteria	Actinobacteria	Actinobacteria	Pseudonocardiales	Pseudonocardiaceae	Amycolatopsis	
bacilli_bacterium_VT13104_LAZH01	Bacteria	Firmicutes	Bacilli	Bacillales	Bacillaceae	Unknown	
bacteroides_fragilis_BOB25_CP011073	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides	
bifidobacterium_adolescentis_150_LBHQ00000000	Bacteria	Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium	
bifidobacterium_angulatum_GT102_LAHN00000000	Bacteria	Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium	
brachymonas_chironomi_DSM19884_ARGE01	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Unknown	
bradyrhizobium_japonicum_FN1_JGCL01	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	Bradyrhizobium	
bradyrhizobium_sp_thb2_AUGA01	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	Bradyrhizobium	
burkholderia_glathei_PML112	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	Burkholderia	
cellulomonas_FA1_LBMY0	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Cellulomonadaceae	Cellulomonas	
cellvibrio_mixtus_J38_ALBT0	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Cellvibrio	
chlamydia_psittaci_HJ_JPIH01	Bacteria	Chlamydiae	Chlamydiia	Chlamydiales	Chlamydiaceae	Chlamydia	
chromobacterium_subtsugae_MWU2920_LCWP0	Bacteria	Proteobacteria	Betaproteobacteria	Neisseriales	Chromobacteriaceae	Chromobacterium	
citrobacter_rodentium_DBS100_JXUN0	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Citrobacter	
clavibacter_michiganensis_DOAB397	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	Clavibacter	
coxiella_burnetii_NLLimburg_JZWL0	Bacteria	Proteobacteria	Gammaproteobacteria	Legionellales	Coxiellaceae	Coxiella	
cronobacter_sakazakii_CDC200903746_JZDO0	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Cronobacter	
deinococcus_sp_RL_JMQF01	Bacteria	Deinococcus-Thermus	Deinococci	Deinococcales	Deinococcaceae	Deinococcus	
delftia_tsuruhatensis_MTQ3_LCZH01	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Delftia	
dickeya_dianthicola_RNS049_APVF01	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Dickeya	
edwardsiella_piscicida_LADL05105_CP011364	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Edwardsiella	
enterobacter_cloacae_UW5_CP011798	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Enterobacter	
enterobacter_sp_54_JFHW01	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Enterobacter	
enterococcus_faecium_ATCC51559_JSVT0	Bacteria	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Enterococcus	
erwinia_tracheiphila_BuffGH_JXNU0	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Erwinia	
erythrobacter_vulgaris_O1_CCSI01	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Erythrobacteraceae	Erythrobacter	
flavobacterium_psychrophilum_FPG3_CP008207	Bacteria	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	Flavobacterium	
flavobacterium_sp_83_JQMS0	Bacteria	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	Flavobacterium	
francisella_tularensis_OR960246	Bacteria	Proteobacteria	Gammaproteobacteria	Thiotrichales	Francisellaceae	Francisella	
frankia_DC12_LANG0	Bacteria	Actinobacteria	Actinobacteria	Frankiales	Frankiaceae	Frankia	
fructobacillus_EFBN1_LDUY01	Bacteria	Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae	Unknown	
gardnerella_vaginalis_3549624_LFWD0	Bacteria	Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Gardnerella	
geobacillus_Zgt1_LDPD01	Bacteria	Firmicutes	Bacilli	Bacillales	Bacillaceae	Geobacillus	
gluconobacter_oxydans_NL71_LCTG0	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Acetobacteraceae	Gluconobacter	
haemophilus_influenzae_MiHi64_JXMG0	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus	
hafnia_paralvei_GTAHAF03_JWGZ01	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Hafnia	
halomonas_lutea_DSM23508_ARKK01	Bacteria	Proteobacteria	Gammaproteobacteria	Oceanospirillales	Halomonadaceae	Halomonas	
halomonas_MCTG39a_JQLV0	Bacteria	Proteobacteria	Gammaproteobacteria	Oceanospirillales	Halomonadaceae	Halomonas	
hassallia_byssoidea_VB512170_JTCM0	Bacteria	Cyanobacteria	NA	Nostocales	Unknown	Unknown	
jiangella_alkaliphila_KCTC19222_LBMC00000000	Bacteria	Actinobacteria	Actinobacteria	Unknown	Unknown	Unknown	
kingella_kingae_KK247_CCJT0	Bacteria	Proteobacteria	Betaproteobacteria	Neisseriales	Neisseriaceae	Kingella	
legionella_neumophila_Bnt314_BBUG0	Bacteria	Proteobacteria	Gammaproteobacteria	Legionellales	Legionellaceae	Legionella	
leuconostoc_mesenteroides_LbE15_JAYN0	Bacteria	Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae	Leuconostoc	
leuconostoc_mesenteroides_P45_JRGZ0	Bacteria	Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae	Leuconostoc	
loktanella_hongkongensis_UST950701009PT_APGJ01	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Loktanella	
lysinibacillus_xylanilyticus_DSM23493_LFXJ01	Bacteria	Firmicutes	Bacilli	Bacillales	Bacillaceae	Lysinibacillus	
magnetospirillum_magnetotacticum_MS1_JXSL01	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Rhodospirillaceae	Magnetospirillum	
mannheimia_haemolytica_D174_CP006574	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Mannheimia	
meiothermus_ruber_A_JXOP01	Bacteria	Deinococcus-Thermus	Deinococci	Thermales	Thermaceae	Meiothermus	
mesorhizobium_loti_NZP2037_AQZP01	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Phyllobacteriaceae	Mesorhizobium	
methylomicrobium_agile_ATCC34068_JPOJ0	Bacteria	Proteobacteria	Gammaproteobacteria	Methylococcales	Methylococcaceae	Methylomicrobium	
micrococcus_sp_MSAsIII49_JXSP0	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Micrococcaceae	Micrococcus	
microvirga_vignae_BR3299	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Unknown	Unknown	
mycobacterium_tuberculosis_134152_LAVG00000000	Bacteria	Actinobacteria	Actinobacteria	Corynebacteriales	Mycobacteriaceae	Mycobacterium	
mycoplasma_gallinaceum_B20968B_CP011021	Bacteria	Tenericutes	Mollicutes	Mycoplasmatales	Mycoplasmataceae	Mycoplasma	
mycrocystis_aeruginosa_NIES2549_CP011304	Bacteria	Cyanobacteria	NA	Chroococcales	NA	Microcystis	
nocardia_seriolae_N2927_BAWD02	Bacteria	Actinobacteria	Actinobacteria	Corynebacteriales	Nocardiaceae	Nocardia	
oenococcus_oeni_139_LCTM0	Bacteria	Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae	Oenococcus	
ornithobacterium_rhinotracheale_ORTUMN88_CP006828	Bacteria	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	Ornithobacterium	
paenibacillus_wulumuqiensis_JCM30284_LAQP01	Bacteria	Firmicutes	Bacilli	Bacillales	Unknown	Unknown	
pantoea_anthophila_112_JXXL01	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Pantoea	
peptococcaceae_CEB3_LDXJ0	Bacteria	Firmicutes	Clostridia	Clostridiales	Unknown	Unknown	
porphyromonas_canoris_COT108_JQZX0	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Unknown	Unknown	
pragia_fortium_24613_CP010423	Bacteria	Proteobacteria	Gammaproteobacteria	Unknown	Unknown	Unknown	
propionibacterium_acnes_ATCC6919_JNHS0	Bacteria	Actinobacteria	Actinobacteria	Propionibacteriales	Propionibacteriaceae	Propionibacterium	
pseudomonas_aeruginosa_WS136_CBXZ00000000	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas	
pseudomonas_syringae_NCPPB2254	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas	
ralstonia_mannitolilytica_MRY140246_BBUP0	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	Ralstonia	
raoultella_terrigena_R1Gly_LANE0	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Klebsiella	
rhodobacter_lobularis_strain_LFTY0	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Unknown	
rhodococcus_sp_IcdP1_CP011341	Bacteria	Actinobacteria	Actinobacteria	Corynebacteriales	Nocardiaceae	Rhodococcus	
rickettsia_hoogstraalii_DSM22243_CCXM01	Bacteria	Proteobacteria	Alphaproteobacteria	Rickettsiales	Rickettsiaceae	Rickettsia	
riemerella_anatipestifer_CH3_CP006649	Bacteria	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	Riemerella	
salmonella_enterica_JOKKOOOOOOOO	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Salmonella	
sar11_bacterium_JPSL01	Bacteria	Proteobacteria	Alphaproteobacteria	Pelagibacterales	Pelagibacteraceae	Candidatus Pelagibacter	
scytonema_millei_VB511283_JTJC0	Bacteria	Cyanobacteria	NA	Nostocales	Scytonemataceae	Unknown	
serratia_liquefaciens_HUMV21_CP011303	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Serratia	
shewanella_sp_ECSMB14012_LWGX01	Bacteria	Proteobacteria	Gammaproteobacteria	Alteromonadales	Shewanellaceae	Bacillus/Shewanella	
sphingomonas_SRS2_LARW01	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingomonas	
staphylococcus_aureus_PK14_CP011528.CP011529	Bacteria	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus	
sulfospirillum_sp_UCH001_AP014723	Bacteria	Proteobacteria	Epsilonproteobacteria	Campylobacterales	Campylobacteraceae	Sulfurospirillum	
thermoanaerobacter_YS13_JOOI01	Bacteria	Firmicutes	Clostridia	Thermoanaerobacterales	Thermoanaerobacteraceae	Thermoanaerobacter	
thermotoga_maritima_Tma200_CP010967	Bacteria	Thermotogae	Thermotogae	Thermotogales	Thermotogaceae	Thermotoga	
thermus_filiformis_ATCC43280_JPSL02	Bacteria	Deinococcus-Thermus	Deinococci	Thermales	Thermaceae	Thermus	
treponema_sp_OMZ838_CP009227	Bacteria	Spirochaetes	Spirochaetia	Spirochaetales	Spirochaetaceae	Treponema	
ureaplasma_diversum_ATCC49782_CP009770	Bacteria	Tenericutes	Mollicutes	Mycoplasmatales	Mycoplasmataceae	Unknown	
vibrio_parahaemolyticus_VH3_LCVL00000000	Bacteria	Proteobacteria	Gammaproteobacteria	Vibrionales	Vibrionaceae	Vibrio	
weissella_ceti_WS08_CP007588	Bacteria	Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae	Weissella	
xanthomonas_sacchari_LMG476_JXQE01	Bacteria	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	Xanthomonas	
xenorhabdus_khoisanae_MCB_LFCV01	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Xenorhabdus	
xylella_fastidiosa_CoDiRo_JUJW0	Bacteria	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	Xylella	

Conclusión del Capítulo 3

La coherencia ecológica de los niveles taxonómicos superiores (dominio, filo, clase, orden, familia y género) es una de las hipótesis más debatidas en el área de la ecología microbiana. Si bien se ha argumentado con diversos ejemplos que los niveles taxonómicos superiores presentan características ecolgógicas distintivas, esta hipótesis no ha sido probada de forma extensiva.

Utilizando los genes codificantes para enzimas presentes en más de 30,000 genomas bacterianos diseñamos modelos basados en algoritmos de aprendizaje supervisado para predecir la pertenencia de cada genoma a los distintos niveles taxonómicos superiores.

La gran precisión alcanzada por estos modelos independientemente del grupo taxonómico observado, utilizando como información de partida a genes directamente involucrados en el metabolismo bacteriano permite trazar una línea de contacto entre la ubicación taxonómica de cada microorganismo y su potencial metabólico codificado en el genoma. Este resultado es el primer argumento a favor de la hipótesis mencionada anteriormente obtenido a partir de un análisis exhaustivo del total de genomas disponibles actualmente.

Adicionalmente, se creó una herramienta que permite inferir la posición taxonómica a niveles superiores a partir de esta información así como explorar los patrones de presencia de genes relacionados al metabolismo y compararlos entre grupos taxonómicos, posibilitando la extracción de información metabólica relevante.

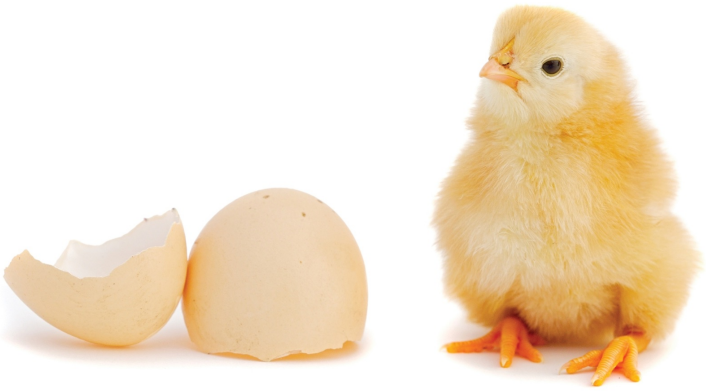
Parte II

Patogenómica del género *Campylobacter*

(Pathogenomics of the genus *Campylobacter*)

4

Campylobacter genomics: emergence of pathogenicity and niche evolution



Citation:

Iraola G*, Pérez R, Naya H, Paolicchi F, Pastor E, Valenzuela S, Calleros L, Velilla A, Hernández M, Morsella C. (2014) **Genomic evidence for the emergence and evolution of pathogenicity and niche preferences in the genus *Campylobacter*.** *Genome Biology and Evolution*. 6(9):2392-2405.

* Corresponding author

4.1 Abstract

The genus *Campylobacter* includes some of the most relevant pathogens for human and animal health; the continuous effort in their characterization has also revealed new species putatively involved in different kind of infections. Nowadays, the available genomic data for the genus comprise a wide variety of species with different pathogenic potential and niche preferences. In this work, we contribute to enlarge this available information presenting the first genome for the species *Campylobacter sputorum* bv. sputorum and use this and the already sequenced organisms to analyze the emergence and evolution of pathogenicity and niche preferences among *Campylobacter* species. We found that campylobacters can be unequivocally distinguished in established and putative pathogens depending on their repertory of virulence genes, which have been horizontally acquired from other bacteria because the nonpathogenic *Campylobacter* ancestor emerged, and posteriorly interchanged between some members of the genus. Additionally, we demonstrated the role of both horizontal gene transfers and diversifying evolution in niche preferences, being able to distinguish genetic features associated to the tropism for oral, genital, and gastrointestinal tissues. In particular, we highlight the role of nonsynonymous evolution of disulphide bond proteins, the invasion antigen B (CiaB), and other secreted proteins in the determination of niche preferences. Our results arise from assessing the previously unmet goal of considering the whole available *Campylobacter* diversity for genome comparisons, unveiling notorious genetic features that could explain particular phenotypes and set the basis for future research in *Campylobacter* biology.

4.2 Introduction

Members of the genus *Campylobacter* are ecologically diverse and naturally inhabit a wide variety of mammals, birds, and reptiles; and some species have an outstanding role in human and animal health [130]. The renewed effort in bacterial characterization driven by the advent and consolidation of next-generation sequencing technologies

has brought back historically underestimated *Campylobacter* species that nowadays pose an emerging risk of zoonotic transmission as they are found in companion, farm and wild animals, and can also contaminate food [131].

The first *Campylobacter* species whose genome was completely sequenced was the foodborne pathogen *Campylobacter jejuni*, the leading cause of gastrointestinal illness in developed and developing countries. This species is subdivided in *C. jejuni* subsp. *jejuni* and *C. jejuni* subsp. *doylei*, which differ in their genomes and pathogenic characteristics [132]. Posterior efforts achieved the sequencing of *Campylobacter coli*, another established pathogen that accounts for 1-25% of *Campylobacter*-related diarrheal diseases [133]. Both *C. jejuni* and *C. coli* are probably the most studied species of the genus and have been subject of continuous research focused on the elucidation of their pathogenic and ecological features from a genomics perspective [134, 135]. The species *Campylobacter upsaliensis* and *Campylobacter lari* complete a list of fully sequenced, gastrointestinal campylobacters that are established human pathogens [136, 137]. Even though most of the species described above historically deserved great attention due to their importance in human health, when the genus *Campylobacter* was first proposed in 1963 it only included two species: *Campylobacter fetus* and "*Campylobacter bubulus*", both isolated from genital tissues of cattle [138]. The species *C. fetus* is now recognized as an established pathogen with a well-known incidence in the reproductive health of cattle and sheep, causing significant socioeconomic burden worldwide [139]. Furthermore, *C. fetus* is subdivided in two subspecies, whose genomes have been also sequenced: *C. fetus* subsp. *venerealis*, a bovine-exclusive pathogen that colonizes the genital tract causing infertility and abortion [140]; and *C. fetus* subsp. *fetus*, which not only causes genital infections in cattle and sheep but also is associated to bacteremia in humans [131]. By the contrary, "*C. bubulus*" has deserved less attention because it is just a commensal bacterium typically found in the genital tissues of healthy cattle [138, 141], although some authors have remarked its role as a putative pathogen causing sporadic infections in humans [142, 143]. To date, *C. bubulus* strains have not been fully sequenced and released. Because the taxonomic structure of the genus has changed extensively, *C. bubulus* was renamed as *Campylobacter sputorum* and divided intraspecifi-

cally in three biovars (bv. sputorum, bv. fecalis, and bv. paraureolyticus), in accordance with their biochemical behavior [144].

Six additional species fill the list of fully sequenced organisms: *Campylobacter concisus*, *Campylobacter curvus*, *Campylobacter rectus*, *Campylobacter hominis*, *Campylobacter showae*, and *Campylobacter gracilis*. As well as *C. sputorum*, these species are remarked as putative pathogens as their prevalences in human or animal infections are widely variable and their clinical presentations are less clear; all of them have been reported at least once causing different kinds of infections [145]. In particular, *C. rectus* and *C. gracilis* have been associated with periodontal diseases and infections in the oral cavity; however, despite a pathogenic role can be suspected for these species robust evidence of causality is still scarce [146]. The species *C. concisus*, *C. curvus*, and *C. showae* have been related to the oral cavity too; nevertheless their role in clinical cases is even less clear.

In summary, the species belonging to the genus *Campylobacter* present a wide phenotypic variability. Based on their pathogenic potential and clinical presentations they can be clearly divided in established and putative pathogens and, although most species can be isolated from different hosts and tissues (niches), they can be grouped in gastrointestinal, oral, and genital depending on their main source of isolation [145]. Nowadays, the number of publicly available genomes for *Campylobacter* species and subspecies are a good representative of the explained phenotypic diversity, allowing to study the relationship between genomic variability, pathogenic potential, and niche preferences. In this work, we report the first completed genome and characterization for *C. sputorum* bv. sputorum strain INTA 08/209, a natural isolation from semen of a healthy bull. We then use this and a representative set of publicly available *Campylobacter* genomes to analyze the emergence and evolution of pathogenicity in this genus using genome-wide comparative analyses and phylogenetics. Furthermore, we focused on the comparison of *Campylobacter* species that are able to colonize different tissues, in order to determine genetic features associated with niche preferences. Our findings suggest that the emergence of pathogenicity can be correlated to the acquisition of virulence genes through horizontal gene transfers from other bacteria and posterior interchange between some members of the genus. Niche preferences can

be mostly explained by nonsynonymous evolution of DSB (disulphide bond) proteins and the invasin CiaB, as well as global compositional differences in GC content, genomes size, and secretomes. In this article, we provide the first comparative genomics analysis of a representative sample of *Campylobacter* taxonomic diversity, pointing the essentiality of sequencing nonclassical organisms to obtain information about the evolutionary mechanisms governing bacterial genomes.

4.3 Methods

4.3.1 Bacterial strains, sequencing and assembly. The strain INTA 08/209 was isolated from the semen of a healthy bull in Argentina in 2008. Samples were inoculated in Skirrow agar and the isolated colonies were classified as *Campylobacter sputorum* bv. sputorum because they were unable to produce catalase and urease, and they were unable to grow in a 1% glycine broth and did produce H_2S. For further confirmation, a fragment of rRNA 16S gene was polymerase chain reaction-amplified using previously described primers and conditions [147] and compared with the 16S ribosomal RNA sequences available in GenBank using BLASTN.

Genomic DNA was isolated with the Wizard Genomic DNA purification kit (Promega), sequencing was performed on an Illumina HiSeq 2000 platform and generated 9,617,780 paired-end reads (2×100 cycles). The resulting reads were first corrected using ALLPATHS-LG [148] and then assembled with Velvet software [149], PAGIT toolkit [150] was used for post-assembly improvement, and the final assembly quality was evaluated with ALE [151]. The resulting contigs were automatically annotated with RAST [152] and manually curated with Pfam and BLASTP over the nr database.

The genomes and available annotations for *C. concisus* 13826, *C. curvus* 525-92, *C. fetus* subsp. *fetus* 82-40, *C. fetus* subsp. *venerealis* NCTC10354, *C. hominis* ATCC BAA-381, *C. jejuni* subsp. *jejuni* RM1221, *C. jejuni* subsp. *jejuni* 55037, *C. jejuni* subsp. *jejuni* 129-258, *C. jejuni* subsp. *jejuni* 51494, *C. jejuni* subsp. *jejuni* LMG9879, *C. jejuni* subsp. *jejuni* LMG9217, *C. jejuni* subsp. *jejuni* LMG23218, *C. jejuni* subsp. *jejuni* 2008-872, *C. jejuni* subsp. *doylei* 269-97, *C. lari* RM2100, *C. gracilis* RM32668, *C. showae* RM3277, *C.rectus* RM3267, *C. upsalien-*

sis RM3195, *C. coli* 76339, *C. coli* BIGS0010, and *C. coli* RM2228 were retrieved from the NCBI. For *C. coli* we considered intra-specific diversity by analyzing representatives from clades 1, 2, and 3 described in Sheppard *et al.* [134]. For *C. jejuni* we included representative genomes from the main clonal complexes described in Sheppard *et al.* [134]. When not available, annotations were generated with RAST. All plots, graphics and data analysis were generated using in-house R scripts.

4.3.2 Orthologous groups, virulence factors, and gene ontologies. The best reciprocal hit approach using BLASTP was implemented to recover shared genes for each pair of genomes, for each reciprocal hit with query coverage >95% and identity >50%. This analysis was complemented running OrthoMCL [153] with default parameters. The set of virulence-associated genes of each genome was recovered using BLASTP against an in-house database created from the virulence factors database [154] and the nr database, and Fisher's exact test was conducted to determine which virulence genes had significant differences among established and putative pathogens ($P < 0.01$).

For functional annotations, each proteome was analyzed with BLAST2GO [155] and gene ontology (GO) terms (belonging to Biological Process) frequency distributions were used to implement the non-parametric Kruskal-Wallis test of variance ($P < 0.01$) in order to identify enriched gene functions among oral versus non-oral, genital versus non-genital, and established versus putative. These groups were defined based on the predominant source of isolation for each species and their pathogenic characteristics (Tab. 4.1).

4.3.3 Phylogenetics, ancestral reconstruction and selection. The consensus phylogeny for *Campylobacter* genomes was obtained from 16S genes aligned with MUSCLE [156] and using Neighbor-Joining and Maximum-Likelihood (ML) methods; and from the full proteome, using an alignment-free method based on protein feature frequency profiles [157]. In order to check if inferred phylogenetic relationships were product of bias in taxonomic sampling, a tree was constructed with 16S genes for all *Campylobacter* species available in SILVA [158] (as author request). In all cases, bootstrap analysis was performed with $1,000$ resampled data sets. To infer phylogenetic re-

Table 4.1: Description of analyzed genomes.

Species	Clade/CC	Accession	Size	GC	Genes	Pathogenicity	Niche
C. coli RM2228	Clade 1	AAFL01	01/01/86	31/01/16	1967	Established	Gastrointestinal
C. coli BIGS0010	Clade 2	ANGU00	01/01/66	31/05/16	1665	Established	Gastrointestinal
C. coli 76339	Clade 3	HG326877	01/01/58	32	1556	Established	Gastrointestinal
C. jejuni 2008-872	61	AIOR00	01/06/16	30/04/16	1702	Established	Gastrointestinal
C. jejuni LMG23218	48	AIOB01	01/06/16	30/04/16	1734	Established	Gastrointestinal
C. jejuni 51494	353	AINZ00	01/08/16	30.2	1975	Established	Gastrointestinal
C. jejuni LMG9217	443	AIOO01	01/06/16	30/03/16	1754	Established	Gastrointestinal
C. jejuni LMG9879	21	AIOI01	01/06/16	30/04/16	1734	Established	Gastrointestinal
C. jejuni 129-258	42	AINY01	01/06/16	30/05/16	1679	Established	Gastrointestinal
C. jejuni 55037	45	AIOH01	01/01/59	30/05/16	1666	Established	Gastrointestinal
C. jejuni jejuni RM1221	354	NC_003912	01/01/78	30/05/16	1838	Established	Gastrointestinal
C. jejuni doylei 269.97	-	NC_009707	01/01/85	30/06/16	1731	Established	Gastrointestinal
C. upsaliensis RM3195	-	AAFJ01	01/01/77	34.2	1934	Established	Gastrointestinal
C. lari RM2100	-	NC_001239	01/01/57	29/06/16	1544	Established	Gastrointestinal
C. fetus fetus 82-40	-	NC_008599	01/01/77	33.3	1719	Established	Genital
C. fetus venerealis NCTC10354	-	AFGH01	01/01/87	33.2	1718	Established	Genital
C. sputorum INTA08/209	-	JMTI0	01/01/78	29	1869	Putative	Genital
C. showae RM3277	-	ACVQ01	02/07/16	45.7	2361	Putative	Oral
C. gracilis RM3268	-	ACYG01	01/02/26	46.6	2847	Putative	Oral
C. hominis ATCC BAA-381	-	NC_009714	01/01/71	31/07/16	1687	Putative	Gastrointestinal
C. rectus RM3267	-	ACFU01	01/02/51	44.8	2971	Putative	Oral
C. concisus 13826	-	NC_009802	02/01/16	39.2	1989	Putative	Oral
C. curvus 525.92	-	NC_009715	01/01/97	44.5	1934	Putative	Oral

lationships from genes selected for further discussion (i.e., *dsbA*), protein sequences were aligned with MUSCLE and ML method was used to build the trees under the Jones-Taylor-Thornton substitutions model. When required, sequences from *Arcobacter butzleri* RM4018, *Arcobacter nitrofigilis* DSM 7299, *Sulfurospirillum barnessi* SES3, and *Sulfurospirillum deleyianum* DSM 6946 were used as outgroups. For each virulence gene, the presence (1) or absence (0) was established as a discrete state in each genome. This information was used to reconstruct the internal nodes states over the consensus phylogeny using the ML method implemented in APE [159].

Screening for diversifying selection over alignment positions was implemented with mixed effects model of evolution with default parameters, due to its ability of detecting episodic selection [160]. For finding conserved positions in alignments among organisms sharing a particular niche preference, we generated in-house R scripts for counting shared positions between them which differ in the rest and created a null distribution counting the same for all possible random groups of genomes.

4.3.4 Secretome analysis. Secreted proteins were predicted for all proteomes using the default settings for Gram-negative bacteria on the

SignalP Server 4.1 [161]. Non-classical secreted proteins for the SignalP were predicted using SecretomeP 2.0 Server [162]. The correspondence analysis between amino acids usage and niche preferences was performed using seqinr [163] and ca [164] packages in R.

4.4 Results and Discussion

4.4.1 ***Campylobacter sputorum*** **GENOME OVERVIEW.** The complete chromosome of *C. sputorum* bv. sputorum strain INTA 08/209 was assembled into 34 contigs of 150-fold in average coverage and 52,394 bp in average length (maximum contig length was 407,694 bp). The estimated chromosome size resulted to be 1,781,420 bp. with an average GC content of 29%. The chromosome contained 1,869 predicted protein-coding genes, 3 rRNA operons, and 42 tRNA genes. The genomic sequences were deposited in the GenBank database under accession number JMTI00000000.

The genome of *C. sputorum* bv. sputorum had the lowest GC content for a *Campylobacter* species reported so far. This feature is probably reflecting a host-associated lifestyle which tends to gradually lower GC content driven by adaptive evolution [165]. *Campylobacter sputorum* bv. sputorum also showed a reduced chromosome size in comparison with other members of the genus and presented the lowest synteny conservation among campylobacters, evidencing that sequence rearrangements have been shaping its genomic architecture, as expected for bacteria which have suffered host-restriction processes [20, 166] (Supp. Fig. 4.1). The species *C. sputorum* bv. sputorum had 181 unique protein-coding genes in comparison with their *Campylobacter* relatives, indicating that the pan-genome of this genus is still open and will probably increase as new complete genomes become available. Moreover, 83 out of these 181 genes exclusive for *C. bubulus* were not found in protein databases (ORFans). Previous analyses have reported that 20-30% of genes present in a novel genome may be ORFans [167], for *C. sputorum* bv. sputorum this percentage is quite lower (4%) probably reflecting a reductional process in the dispensable genes set for this species. The SAP (surface array protein) genes, which are the main antigenic determinants for the genital species *C. fetus*, were not present in *C. sputorum*. The genome of *C. sputorum* bv. sputo-

rum presented no evidence for textrachromosomal replicons. Nevertheless, signatures for horizontal gene transfers were identified as some genes were found on other bacterial genomes whereas absent in campylobacters, for example, one gene coding for a putative hemolysin was probably acquired from *Wolinella succinogenes* and suffered posterior pseudogenization by nonsense mutations, showing that this kind of evolutionary process has also been shaping the genomic landscape of *C. sputorum* bv. sputorum. Additionally, two contiguous genes coding for an AAA+ ATPase and a restriction endonuclease were probably acquired from the gram-positive coccus *Eremococcus coleocola*, a relatively new species originally isolated from vaginal tissues from horses [168]. This result provides strong evidence for the recent acquirement of new genes from bacteria that share the niche with *C. sputorum* bv. sputorum, even being phylogenetically distant. With the incorporation of this newly sequenced species we recalculated the core genome for *Campylobacter* genus, estimated in 669 genes.

4.4.2 Evolution of pathogenicity. To study how the pathogenic character has evolved along the *Campylobacter* taxonomy we used 23 available genomes predefined in two groups, considering their documented clinical incidence: 1) *C. coli* strains, *C. jejuni* subsp. *jejuni* strains, *C. jejuni* subsp. *doylei*, *C. lari*, *C. upsaliensis*, *C. fetus* subsp. *fetus*, and *C. fetus* subsp. *venerealis* represent established pathogens in human and/or cattle, and 2) *C. sputorum*, *C. curvus*, *C. concisus*, *C. hominis*, *C. rectus*, *C. showae*, and *C. gracilis* represent putative pathogens. Based on this classification, we looked for the presence or absence of virulence genes on each genome and constructed a presence/absence matrix considering 255 different genes belonging to the following functional categories: Capsular, general glycosylation pathway, SAP, CDT (cytolethal distending toxin), transcription, chemotaxis, plasmid partitioning, outer membrane protein, invasion, adhesion, LPS (lipopolysaccharide), and motility. Fig. 4.1A shows the number of virulence genes identified per genome displayed in accordance with a consensus phylogeny (see Methods). Established pathogens posed a significantly wider repertory of virulence genes ($P = 0.0002$, Fisher's exact test) which also correlated with their phylogenetic position. The unique exception was for *C. fetus* subspecies, that are phy-

logenetically closer to putative pathogens but have an expanded repertory of virulence genes as expected for established pathogens. Taking into account the number of genes belonging to each functional category, species clustered in two distinctive groups that match perfectly with the predefined established and putative pathogens (Fig. 4.1B). Established pathogens were richer in genes coding for LPS, adhesion and, motility, although, the presence of capsular genes (SAP for *C. fetus* subspecies) and CDT were the most relevant features that distinguished established from putative pathogens (P = 0.001, Fisher's exact test). In particular, some authors have questioned the role of CDT as a virulence factor because some naturally occurring *C. jejuni* strains presented partial disruption or absence of CDT operon. As a complementary approach, we screened 85 additional *C. jejuni* genomes and found a prevalence of 90% for *cdtA*, 97% for *cdtB* (the main toxin component), and 98% for *cdtC*. The analysis of several publications [169–172] screening CDT genes in clinical cases or *C. jejuni* populations also showed a prevalence higher than 95%. Moreover, the insertional inactivation or complete deletion of *C. jejuni* CDT genes has been demonstrated to cause reduced invasiveness and adherence, attenuation and asymptomatic infections [173, 174]. These results support the role of CDT as a virulence factor, which probably has a complementary activity with other virulence factors like motility proteins, adhesions, and invasins. However, the occurrence of CDT-negative strains isolated from clinical cases remains as an open question, in spite of being the vast minority. In summary, these results point that pathogenic potential of *Campylobacter* species may be correlated with the presence of certain virulence genes. However, considering the extreme complexity in defining bacterial pathogenicity these results are useful for suggesting general differences among established and putative pathogens, whereas deeper analyses involving experimental approaches should be conducted in the future to decipher the role of these genes during infection.

In order to elucidate the most probable evolutionary path that lead to the actual distribution of these genes among taxa, we implemented an ancestral character reconstruction approach using the presence/absence of virulence genes as states. In first place, the MRCA (most recent common ancestor) for was probably a non-pathogenic (genes presence

probability <0.5) (Fig. 4.2A). However, some virulence genes, such as the invasion antigen B (CiaB), were also found in the *Campylobacter* ancestor, which suggests that ancient campylobacters had the potential to invade host cells. The relevance of CiaB in pathogenic phenotypes will be further discussed afterwards.

The MRCA for putative pathogens showed a probability distribution for genes presence that resembled the *Campylobacter* MRCA (Fig. 4.2B), indicating that the reduced amount of virulence genes present in putative pathogens were present in the ancestor or have been probably acquired through recent horizontal gene transfer events. On the contrary, the MRCA for established pathogens already carried genes for CDT, capsule, motility, LPS, and adhesion (P >0.9), suggesting that these organisms have evolved from an ancestor with a significant virulence armament, acquired from other bacteria (Fig. 4.2C). It is worth mentioning that no significant differences were found among *C. coli* strains belonging to clades 1, 2, and 3. Recent works have shown that *C. coli* clade 1 (the most frequently isolated from clinical cases) has suffered a progressive genomic introgression with *C. jejuni*, whereas clades 2 and 3 are mainly constituted by non-introgressed isolates from riparian environments [134]. However, strains belonging to clades 2 and 3 have been also found in clinical cases and have genes associated with [135]. Based on these results, it is probable that *C. coli* strains have evolved from the same pathogenic ancestor and shaped their genomes for environmental diversification while conserving genes for CDT, capsule, motility, LPS, or adhesion, being currently underreported in clinical cases due to niche separation. The analysis of *C. jejuni* strains belonging to different clonal complexes also revealed no significant differences in their repertories of virulence genes, showing that intraspecific diversification may be linked to the evolution of different genomic components.

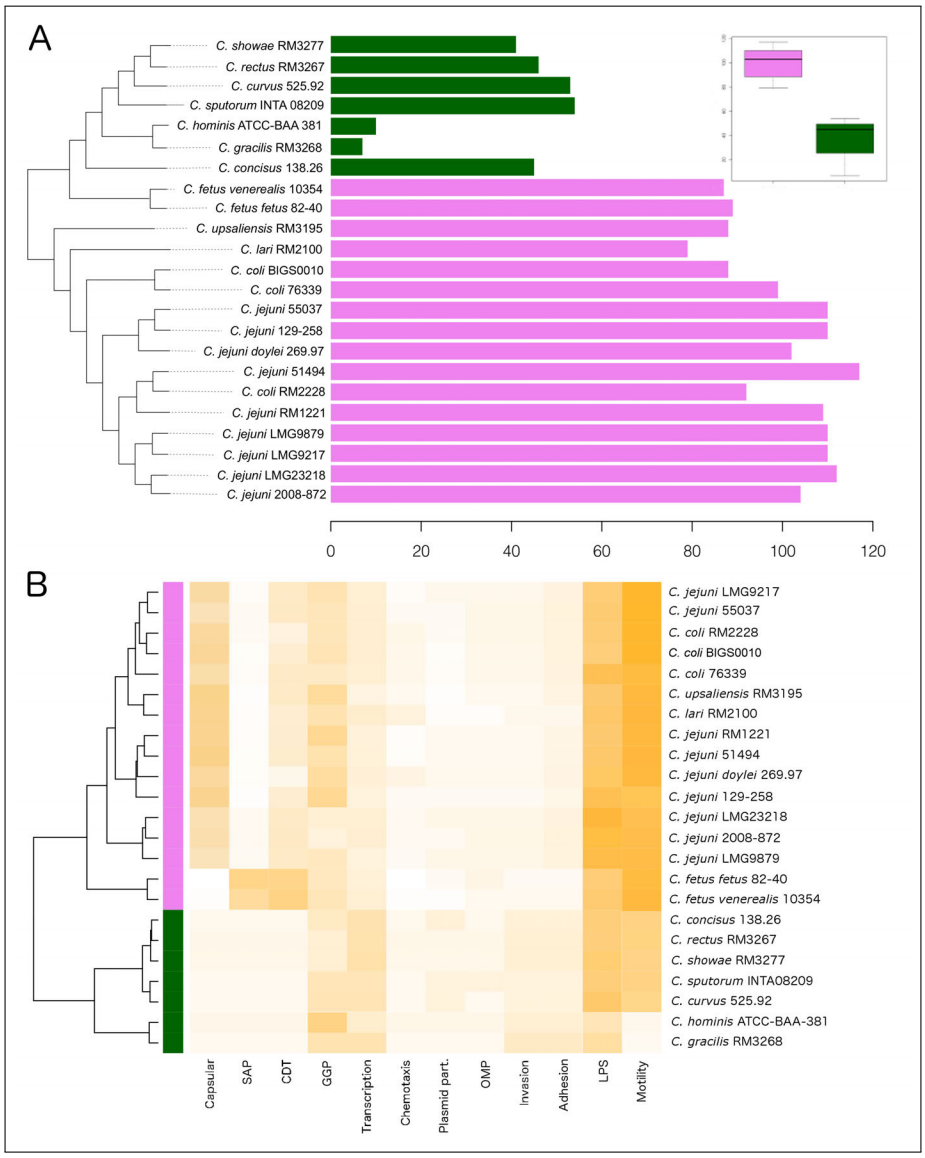


Figure 4.1: Barplot and heatmap of virulence genes identified per genome. Established pathogens are displayed in violet, and putative pathogens are displayed in dark green. (A) The total number of virulence genes is displayed as bar lengths. Genomes are clustered based on the inferred consensus phylogeny for *Campylobacter* genus. (B) Genomes are clustered in established (top) and putative (bottom) pathogens based on the presence/absence patterns for virulence genes belonging to 12 functional categories. Colors (white to orange) represent the number of genes.

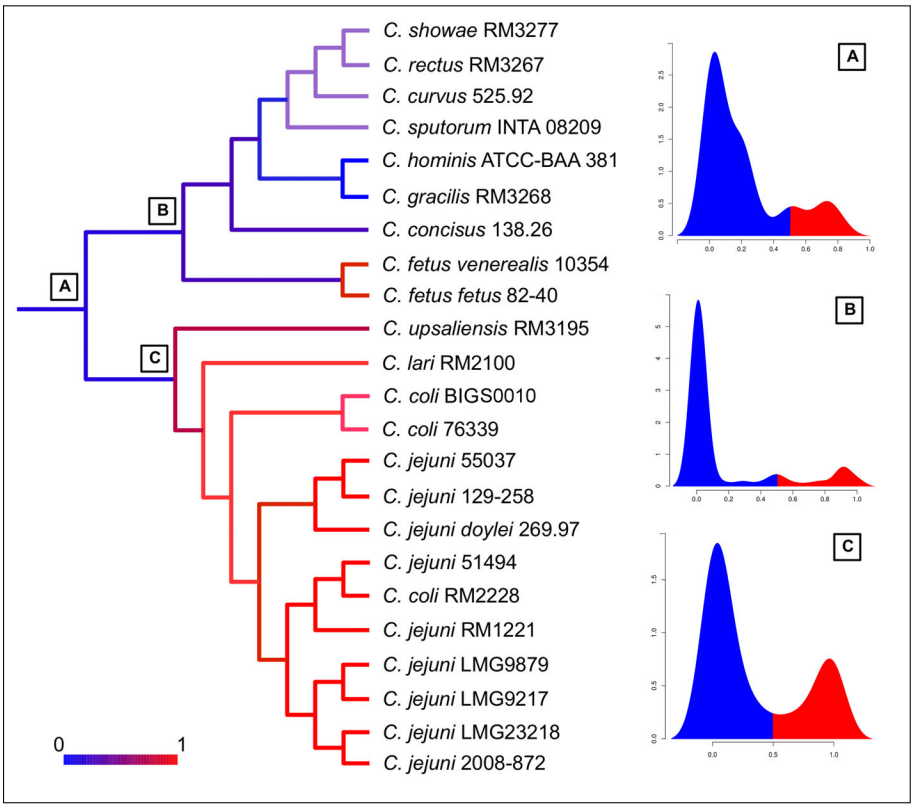


Figure 4.2: Ancestral character reconstruction. The consensus phylogeny for *Campylobacter* is colored according to the average probability for the absence (P = 0, blue) or the presence (P = 1, red) of virulence genes. The probability densities using the inferred states for each genes are shown for the *Campylobacter* MRCA (A), the MRCA for putative pathogens (B), and the MCRA for established pathogens (C).

Notoriously, *C. fetus* subspecies should be classified as established pathogens based on their virulence genes repertories and clinical presentations; however, they are phylogenetically closer to putative pathogens. The MRCA for *C. fetus* subspecies showed the presence of almost all genes found in the MRCA for established pathogens, suggesting that *C. fetus* subspecies evolved from a non-pathogenic ancestor shared with putative pathogens, but acquired a set of virulence genes from species belonging to established pathogens. This kind of horizontal evolution has been documented through plasmid transfer from *C. jejuni* to *C. fetus*, moreover *C. fetus* genomes show extensive evidence of recent horizontal gene transfer events [175, 176]. The lack of a genome

for a sister species to *C. fetus*, like *Campylobacter hyointestinalis*, prevent more accurate estimates for the acquirement of these genes.

In summary, from the analysis of this set of genomes we can conclude that the most probable scenario for the evolution of pathogenicity in *Campylobacter* is the accumulation of virulence factors that resulted in established pathogens, instead of an opposite scenario of pathogenicity attenuation by gene loss from a virulent ancestor. However, posterior gene loss events among putative pathogens should not be discarded, especially for *C. hominis*, which present the smallest virulence armament. Not surprisingly, it was originally isolated from healthy humans [177] and has the lowest number of reported infections among sequenced species. Probably, a more complete representation of *Campylobacter* species could help to better understand the dynamics of horizontal gene transfers and their implications in pathogenicity.

4.4.3 Comparative functional analysis. Beyond the set of virulence genes analyzed so far are part of the best-known players in pathogenic phenotypes, the presence, absence or enrichment in other kind of virulence-associated or virulence lifestyle genes (typically coding for more general metabolic pathways) may be directly implied in pathogenicity [178]. For this reason, we performed a comparative functional analysis of *Campylobacter* proteomes based on GO terms. This approach was also useful to have a first glance of the main metabolic functions associated to niche preferences (Fig. 4.3).

Established pathogens were enriched in functions related to antibiotic resistance (GO:0046677). This feature is particularly interesting because of the central role of antibiotics in the treatment of bacterial infections. The most distinctive feature was the presence of a gene coding for the enzyme aminoglycoside n3-acetyltransferase among established pathogens, whereas absent in all putative pathogens. This enzyme is involved in the resistance to aminoglycosides, which has been extensively proved for *C. jejuni* and *C. coli* [179]. The GO analysis also revealed that established pathogens were enriched in terms related to adhesion (GO:0007155) and motility (GO:0040011), in accordance with the results described in the previous section.

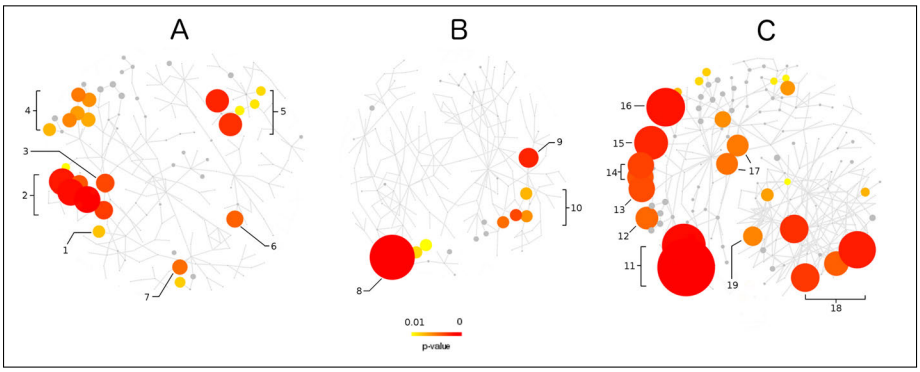


Figure 4.3: GO analysis. Three GO graphs for (A) oral versus non-oral, (B) genital versus non-genital, and (C) established versus putative. Significant GO terms (P <0.01) for each graph are colored in a yellow to red gradient. Numbers encode the name of significant functional categories, (A) 1-pathogenesis, 2-response to virus, 3-proteolysis, 4-response to oxygen species, 5-lactate metabolism, 6-sulfate metabolism, 7-antibiotic resistance; (B) 8-protein methylation, 9-response to external stimulus, 10-nitrogen utilization; (C) 11-response to antibiotics, 12-locomotion, 13-oxydative stress, 14-starvation, 15-nitrogen transport, 16-chromosome partition, 17-adhesion, 18-vitamin biosynthesis, and 19-pathogenesis.

In contrast with putative pathogens, all species belonging to established pathogens are able to efficiently invade host cells [180, 181]. When bacteria invade eukaryotic cells, they are immediately exposed to unfavorable conditions mainly associated to starvation (nutrients shortage) and multiple types of stresses, most notably oxidative stress [182]. Seven genes (coding for thiorredoxin reductase, thiol peroxidase, catalase, methionine sulfoxide reductase, superoxide dismutase, HMBPP reductase, and carbon starvation protein) were involved in the response to starvation and oxidative stress (GO:0006979 and GO:0042594, respectively) and were significantly (P = 0.005) more abundant in established pathogens (Fig. 4.4) evidencing that these metabolic functions could be linked to the pathogenic potential.

A particularly relevant feature for species that can colonize genital tissues was the enrichment in genes for nitrogen metabolism (GO:0019740), which is an integrated mechanism that detects the depletion of the primary nitrogen source and activates genes for scavenging and transporting alternative nitrogen sources. There is scarce information about abundance of nutrients in the genital and urogenital tissues; however, it has been demonstrated that uropathogenic *Escherichia*

coli strains need to activate nitrogen utilization pathways during colonization of mice urinary tract [183]. Other genital bacteria, like *Gardnerella vaginalis*, also encode genes important for the utilization of various nitrogen sources [184] and the pathogen *Candida albicans* (despite non-bacterial) up regulates genes involved in nitrogen utilization when infecting genital tissues [185]. These results indicate a possible role of nitrogen metabolism on the establishment of microorganisms in the apparently hostile genital environment. Genital campylobacters were also enriched in genes involved in protein methylation (GO:0006479). In general, methylation is involved in cell-environment interactions; however, this characteristic needs to be further investigated in order to establish its relation to genital tropism.

Oral campylobacters are suspected pathogens in periodontal diseases, often presenting a complex etiology mainly attributed to polymicrobial disruption of host homeostasis [186]. Recently, the presence of sulfate-reducing bacteria in the complex oral flora has been proposed as implicated in the development of periodontal diseases [187]. Accordingly, oral campylobacters resulted to be enriched in functions related to sulfate metabolism (GO:0006790). These species were also enriched in genes for lactate metabolism (GO:0006089), which plays an important role in the development and maintenance of acidic conditions in vivo. Microbial flora present in cariogenic plaques produce lactate as the predominant glucose-derived product, which is considered to be the main acid involved in caries formation [188]. Because dynamics of periodontal infections are complex, and beyond their direct incidence on oral diseases, the capacity of these *Campylobacter* species to produce lactate may be contributing to the development and establishment of infections, as other microorganisms directly associated to periodontal diseases (like *Streptococcus* or *Veillonella*) are benefited by this lactate-rich environment [189].

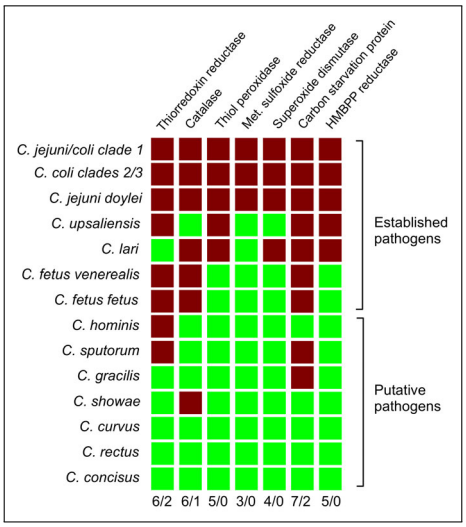


Figure 4.4: Distribution of genes belonging to oxidative stress and starvation. Red boxes show the presence of a gene in a certain genome whereas green boxes show its absence. Fractions at the bottom represent the counting of each gene in established and putative pathogens, respectively.

4.4.4 Secretomes, compositional differences and selection. To gain more insights on the mechanisms involved in niche preferences among *Campylobacter* species, we centered our attention on their predicted secretomes and the differential amino acids usage within the whole proteomes and the secretomes. Proteins with secretory signals are the main tools that bacteria use to interact with their environments [190], so secretome evolution may be driven by host-microorganism interactions which are determined by different types of tissue-specific molecules and environmental conditions. Extensive bioinformatics comparative studies of bacterial secretomes have suggested that secretome size is not correlated with pathogenic potential nor niche preferences at highest taxonomic levels [191]. Among *Campylobacter* genomes we found great differences in predicted secretome sizes, ranging from 80 proteins in *C. sputorum* to 210 in *C. gracilis*. Furthermore, when exploring the number of secreted proteins normalized by the species proteome size, we found that oral species secrete around 10% of their proteins, whereas non-oral species secrete around 6% (Supp. Fig. 4.2). These results indicate that particular, oral-exclusive secreted

proteins should be playing an important role in niche preference for oral cavity and do not support the findings reported by [191]; however, this discrepancy could be explained because the observed signal can be stronger for particular organisms at lower taxonomic levels. Among these oral-exclusive secreted proteins it is worth noting the presence of a divergent kind of metal scavenging TonB-dependent siderophore transporter (TBDT). Metal ions are essential cofactors needed for the correct functioning of most enzymes and bacteria have evolved special macromolecular mechanisms to sequester them from the environment when lacking [192]. TonB systems are formed by an energy transduction complex (TonB-ExbB-ExbD) anchored to the inner plasma membrane and a pore-forming TBDT anchored to the outer membrane [193]. These systems have been well-characterized in *C. jejuni* and *C. coli* to a lesser extent, showing redundancy. Here, we found that genes coding for TonB-ExbB-ExbD complex were conserved in all genomes and showed great synteny conservation too. When exploring TBDT genes we found great diversity in sequence identity and copy number (up to 21 in *C. curvus*). Fig. 4.5 shows the phylogenetic relationships among 98 recovered orthologs, highlighting the presence of different TBDT types like CirA, FuhE, CfrA, and an uncharacterized oral-exclusive cluster. In terms of genomic context, these oral-exclusive TBDTs were in proximity with genes coding for ATP-binding, permease and periplasmic proteins belonging to iron ABC transporters; a methlytransferase domain protein was habitually found next to the TBDT gene too. In *C. showae* and *C. rectus* we found two adjacent TBDT copies probably generated by paralogy, indeed, in *C. showae*, one of them was pseudogenized by nonsense mutations. Additionally, in all cases the genomic surroundings were rich in small predicted hypothetical proteins, suggesting that these TBDTs are placed in plastic regions suffering rearrangements and horizontal transfer events, as members of this divergent cluster were not found in other campylobacters and presented less than 25% of identity with the rest of TBDTs recovered from *Campylobacter* genomes. No significant differences in secretome size and composition were observed for genital and gastrointestinal campylobacters, suggesting that niche preferences do not depend on the evolution of the same set of genes for adaptation to different environments.

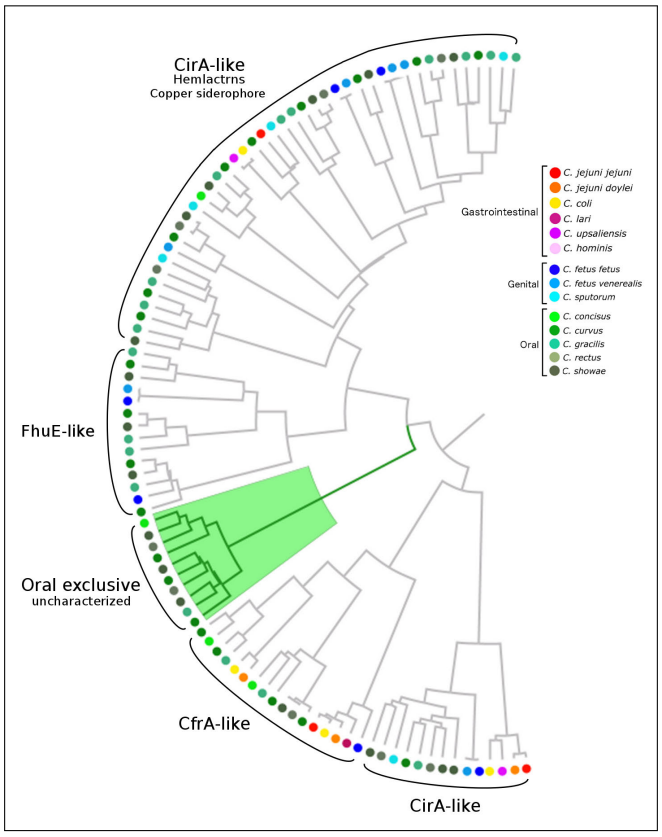


Figure 4.5: Phylogeny of TBDTs. The phylogenetic tree was constructed using 98 TBDT orthologs recovered from *Campylobacter* genomes. The oral-exclusive cluster is highlighted in dark green.

When considering the amino acids usage for secreted proteins, we identified significant differences between *Campylobacter* species belonging to different niches. The correspondence analysis displayed in Fig. 4.6A demonstrated how amino acids usage clearly discriminates oral, genital, and gastrointestinal species in distinctive clusters. The unique exception was for *C. hominis*, which clustered closer to genital species despite being gastrointestinal; this species carried an extremely reduced virulence genes repertory and posed the lowest number of documented infections. It is probable that these particular features are also being reflected in this discrepancy and further investigation is needed for elucidating why the observed amino acids usage was not correlated with the phenotype of this neglected species. Despite

the variability in genome and proteome sizes, the behavior observed in Fig. 4.6A is maintained when using the whole proteome for amino acids usage calculations, suggesting global reach patterns that link non-synonymous evolution with niche preferences. The comparison of amino acids usage from different bacteria has showed that adaptive pressures over amino acids are highly variable along taxonomy [194]; however, correlations between amino acids usage and niche preference or tissue tropism have been proposed only for viruses [195]. We suggest that differences found in amino acids usage among *Campylobacter* species may be attributed to adaptive evolution driven by niche-specific environmental conditions. This is also evident when analyzing the global GC content and genome sizes with respect to different niche preferences, especially for oral species which were distinguished by bigger genomes and higher values of GC content (Fig. 4.6B).

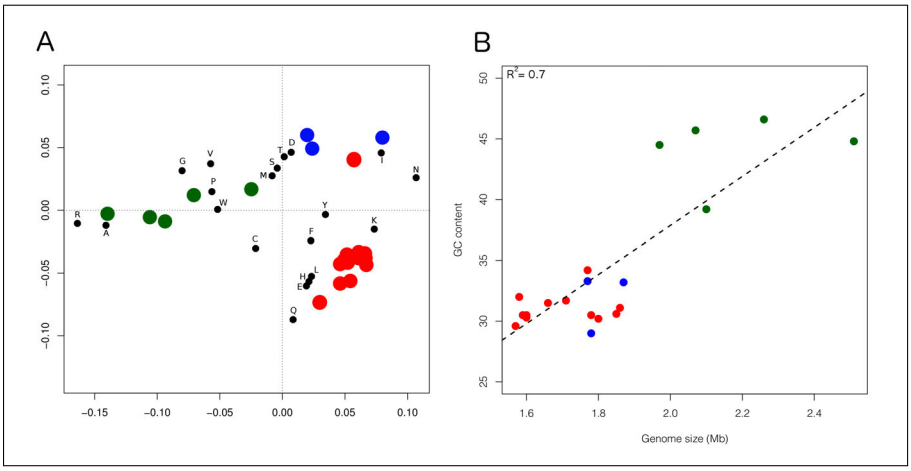


Figure 4.6: Correspondence analysis and whole-genome compositional features. Correspondence analysis using amino acids usage form secreted proteins (A) and linear correlation for genome size versus GC content (B). Small black circles represent each amino acid using the one-letter code. Big circles represent each genome colored according to niche preferences: gastrointestinal (red), genital (blue), and oral (green).

In order to further investigate the correlation between secretome evolution and niche preferences we analyzed the genetic variability among DSB proteins. The DSB system (essentially conformed by *dsbA* and *dsbB* genes) is involved in imparting structural stability to proteins by catalyzing the oxidation of cysteine residues to form DSBs and

is particularly important for the correct folding of secreted proteins. The high sequence variability found among bacterial DSBs indicates that they probably have different substrate specificities [196]; hence, the presence of divergent sets of DSB systems may be linked to the observed differences in *Campylobacter* secretomes. DSB orthologs present in the 23 *Campylobacter* genomes and related genera (*Sulfurospirillum* and *Arcobacter*) were recovered using BLAST searches against annotated DSB genes and analyzed using phylogenies. The main component of DSB system (*dsbA*) was found in all genomes and copy number varied from 1 to 3 (Supp. Tab. 4.1). The phylogenetic reconstruction using *dsbA* orthologs showed the presence of different groups that correlated with niche preferences and evidenced a great deal of gene duplication. Fig. 4.7 shows that groups 1 and 2 share a recent common ancestor and are formed by the same oral *Campylobacter* species, indicating recent paralogy for this divergent set of DSBs probably associated to niche preference for the oral cavity. Group 3 is just composed by the unique organisms capable of colonizing genital tissues (*C. fetus* subspecies and *C. sputorum*), reinforcing the hypothesis of niche-driven evolution of DSB proteins. This theory is additionally supported by the configuration of groups 4 and 5, exclusively formed by established pathogens (which are gastrointestinal).The ancestral genera *Sulfurospirillum* and *Arcobacter* clustered together (group 6), denoting an ancestral vertical evolution of DSB systems among these taxa. In addition, these species showed the lowest gene diversification level, suggesting that DSB systems have experienced a duplication boost since *Campylobacter* diverged, with posterior specialization. Finally, this scenario is similar for DsbB protein, denoting the coevolution of this pair of functionally related genes (data not shown).We also found orthologs for *dsbD* and *dsbE* in the ancestral genera and in some *Campylobacter* species, suggesting gene loss events during the evolution of this genus, probably due to the nonessential functions of these genes.

4.4.5 Host cells invasion and adhesion. The ability of *Campylobacter* species to invade host cells is a well-recognized virulence mechanism in all the established pathogens [180, 181, 197] and in some putative pathogens like *C. concisus* [198]. This phenotype is strongly

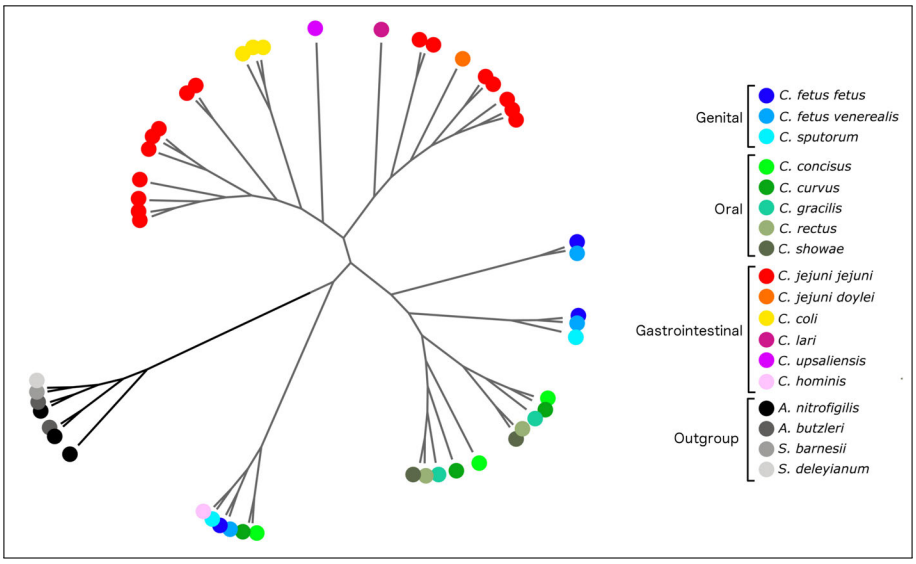


Figure 4.7: Phylogeny of DsbA. The phylogenetic tree using DsbA protein clusters *Campylobacter* and related genomes according to their niche preferences.

correlated with the presence of the invasion antigen B (encoded by *ciaB* gene) which is the main genetic determinant for invasiveness in *Campylobacter*. In this study, we found single copy orthologs for this gene in all the species (average amino acid identity of 70%), even in *C. showae* and *C. hominis*, whose invasive capacity is apparently null [198]. These results open two alternative hypotheses: 1) just the presence of *ciaB* is not enough to warrant a successful invasion, considering that both *C. showae* and *C. hominis* presented a reduced repertory of virulence genes; or 2) sequence variation at amino acid level was responsible for function switching and/or specialization of this gene. In this sense, signals for diversifying selection were found on 34 over 607 (~5%) codon positions in the *ciaB* alignment Fig. 4.3). Furthermore, we looked for shared positions within species belonging to the same niche in order to explain diversifying evolution as function of niche pressures. For *C. jejuni*, *C. coli*, *C. lari*, and *C. upsaliensis* (gastrointestinal established pathogens) we found 24 conserved positions that carried any different amino acid in the rest of the species, suggesting a strong diversifying pressure acting over *ciaB* gene for these phylogenetically related organisms. The scenario for genital species is slightly different

because *C. fetus* subspecies are closely related but *C. sputorum* is phylogenetically distant, even though they shared 11 conserved positions that were different from the rest (three of them also showed positive selection signal), being the maximum number found for any possible species trio and significantly departing from the null distribution (see Methods for details). These results suggest that coevolution has been acting over these sites and reflect the probable specialization of CiaB to invade genital tissues. No associations were found for oral species.

Surface attachment to host cells is the previous step required for invasion. We previously described that genes coding for adhesins were overrepresented in established pathogens, evidencing that some adhesins are exclusive for these organisms. Although, we also found one gene coding for a fibronectin/fibrinogen-binding protein with ubiquitous distribution among campylobacters, suggesting that all species pose a basal attachment potential. The analysis of this gene showed the presence of 25 over 444 (~6%) sites under diversifying selection and 21 sites conserved among gastrointestinal pathogens whereas different in the rest. For this gene, no significant differences were found for genital or oral species. The role of diversifying selection has been previously highlighted for some *Campylobacter* genomes [199], here we show how this evolutionary force is acting in some relevant genes and is probably driven by the particular environmental conditions found in different niches.

4.4.6 The evolutionary mechanism of *Campylobacter* pathogenicity. The whole set of results obtained in this work allow us to accommodate an integrative hypothesis about *Campylobacter* evolution in terms of pathogenicity and niche preferences. Fig. 4.8 shows a summary of the main forces shaping the evolutionary landscape of this genus. On one hand, horizontal gene transfers were probably the main evolutionary force involved in the emergence of some *Campylobacter* species as established pathogens. Probably the group of species conformed by *C. jejuni* (both subspecies) *C. coli*, *C. lari*, and *C. upsaliensis* gradually acquired a set of virulence genes from other bacteria and then transferred most of them to *C. fetus* subspecies. Based on this, pathogenic potential (established or putative pathogen) can be correlated with the presence/absence of certain genes with previous associ-

ation to *Campylobacter* virulence (like CDT, capsule, or flagellum). On the other hand, gene diversification seemed to be playing a central role in the adaptation of species to different environmental conditions. This was suggested from the diversifying evolution of DSB orthologs, CiaB, and the whole secretome. However, the role of horizontal gene transfer events in niche preferences should not be discarded, because we found evidences for the acquisition of genes coding for secreted proteins in oral campylobacters and foreign genes in *C. sputorum* probably acquired from other non-*Campylobacter* genital species.

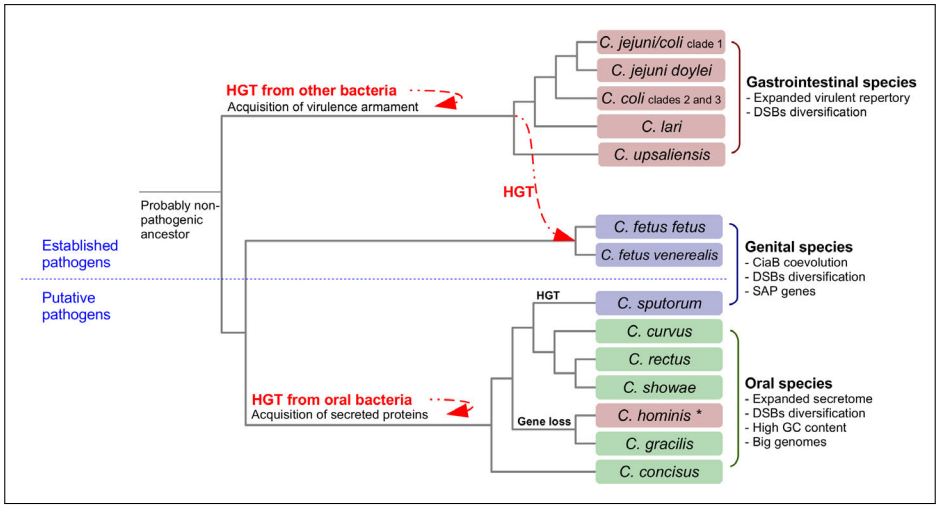


Figure 4.8: Main evolutionary processes in *Campylobacter*. This figure provides a phylogeny-based integrative view of the main evolutionary processes that have been shaping *Campylobacter* genomes in terms of pathogenicity and niche preferences. Species are highlighted in blue (genital), red (gastrointestinal), and green (oral). The species *C. hominis* is ticked off for not sharing the same genomic features than oral species, despite of belonging to the same phylogenetic group.

Despite the evolution of pathogenic potential and niche preferences should be somehow related, the presence of both established and putative pathogens colonizing the genital and gastrointestinal tracts indicated that they are not completely linked. This opens new questions regarding the relationship between niche and virulence, suggesting that genetic features that determine these phenotypes have different patterns of evolution. The following example involving genital species clearly describes this situation: The genome of *C. sputorum* did not

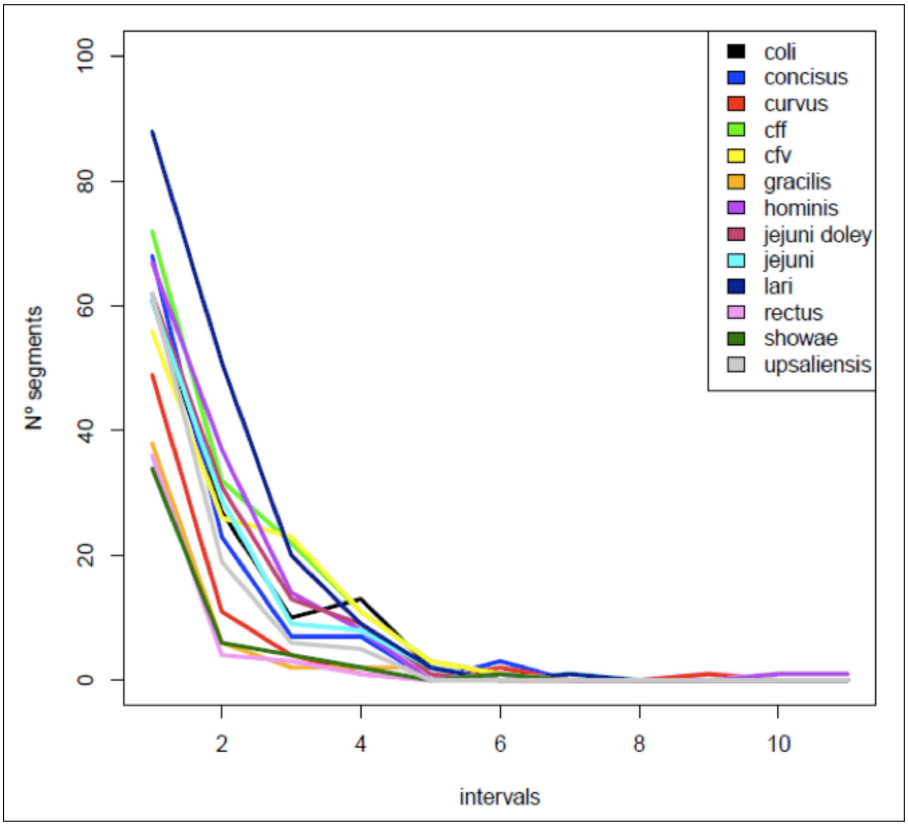
present genes coding for SAP, which are the main antigenic determinants in *C. fetus* and have been associated to virulence in the context of genital infections [200, 201]. On one hand, *C. sputorum* is capable of colonizing genital tissues without causing disease and did not code for SAP, suggesting that these proteins would not be essential for genital tropism. On the other hand, *C. fetus* is capable of causing infection in genital tissues and codes for SAP, so we propose that these proteins should have a role in virulence once the bacterium has been established in the tissue, more than in determining niche preference. How SAP genes emerged in *C. fetus* and why they were not transferred to other species, as well as other virulence genes, is an open question whose answer will involve a detailed study of horizontal gene transfer mechanisms in the context of *Campylobacter* infections.

Finally, the possibility of developing integrative comparative genomics analyses oriented to associate particular genomic features and evolutionary processes to phenotypes, not only depends on the availability of significant species for human health, but also in obtaining genomic information from neglected or less glamorous organisms. Additionally, the best scenario for performing these kinds of analyses should include many representative genomes form each species. The results presented here are product of comparing single representatives for certain campylobacters, evidently not considering the possible intra-specific variability of these species. However, this limitation could be improved in the near future as new genomes become available for different strains of the same species. This could provide the genetic information needed for refining our results and for gathering further genomic evidences for the evolution of pathogenicity and niche preferences among *Campylobacter* species.

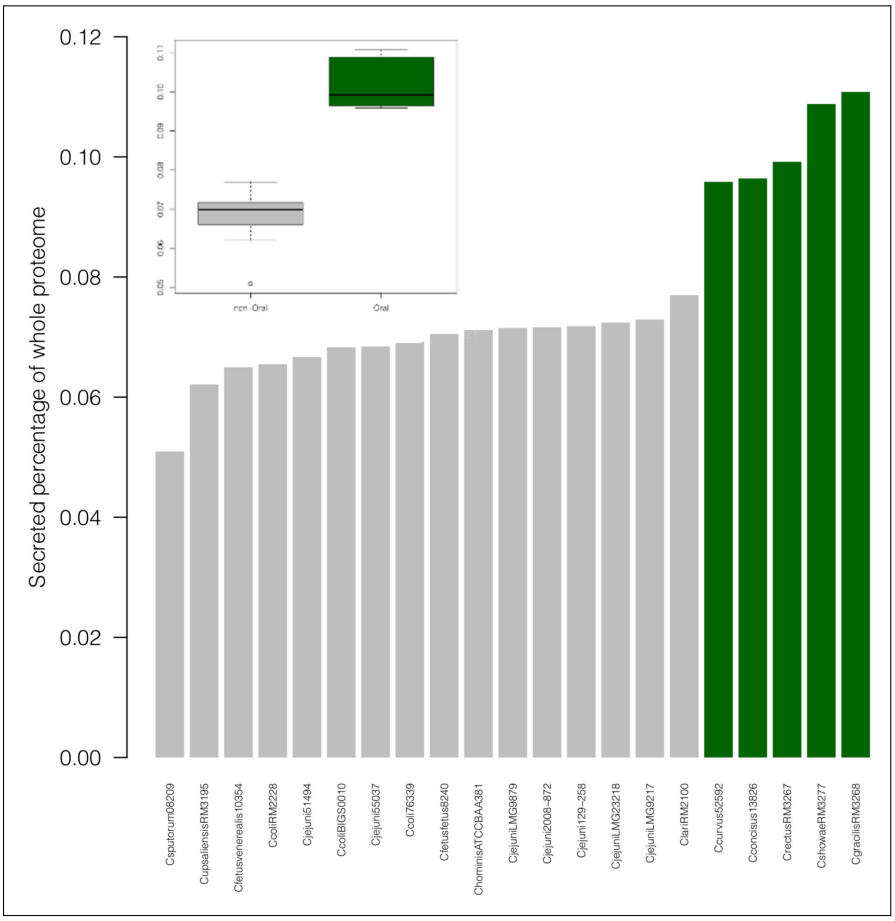
4.5 Acknowledgments

The authors thank Trevor D. Lawley and David Harris from The Wellcome Trust Sanger Institute for invaluable assistance during genome sequencing process. The authors also thank Nicolás Sarute and Rodrigo Martino for reading the manuscript and providing constructive comments.

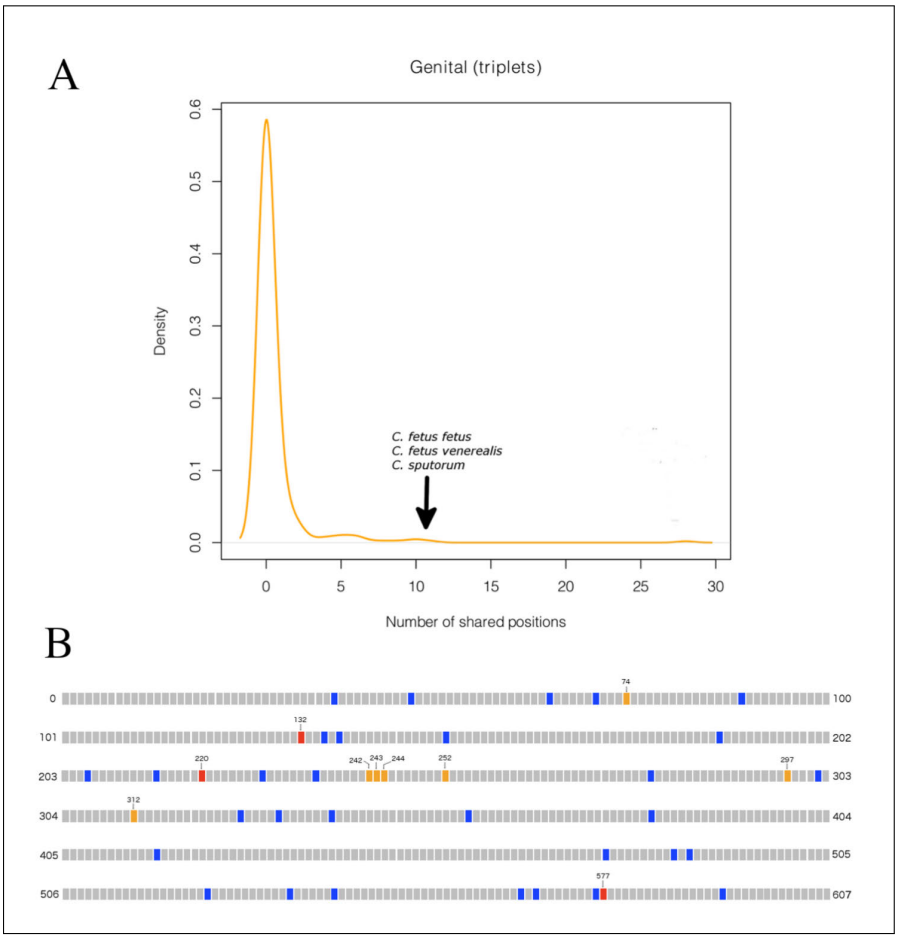
4.6 Supplementary material



Supp. Fig. 4.1: Synteny analysis. The plot shows the frequency of genomic fragments (intervals) up to 10.000 bp. shared between *C. sputorum* and the remaining genomes analyzed. The genomes of *C. jejuni* RM1221 and *C. coli* RM2228 were used in this analysis.



Supp. Fig. 4.2: Secretome sizes. Each bar shows for each genome, the percentage of its total proteome that was predicted as secreted. Green bars belong to oral campylobacters. The boxplot at the top is an alternative presentation of the same data, highlighting the significant differences between oral species and the rest.



Supp. Fig. 4.3: Selected positions over *ciaB* gene. Distribution of conserved positions in *ciaB* alignment for all possible triplets of genomes (A). Positions correspond to each codon/amino acid in the alignment of *ciaB* orthologs. Blue positions have signal for diversifying selection. Orange positions are shared among genital species while different in the rest. Red positions satisfy both conditions (B).

Supp. Tab. 4.1: Number of genes coding for DSB proteins among genomes.

	dsbA	*dsbB*	*dsbD*	*dsbE*
C. coli RM2228	1	1	1	1
C. coli BIGS0010	1	1	1	1
C. coli 76339	1	1	1	1
C. jejuni 55037	2	1	1	2
C. jejuni 129-258	2	1	1	2
C. jejuni 51494	2	1	1	2
C. jejuni RM1221	2	1	1	2
C. jejuni LMG9879	1	1	1	2
C. jejuni LMG9217	2	1	1	2
C. jejuni LMG23218	2	1	1	2
C. jejuni 2008-872	2	1	1	2
C. jejuni doylei 269.97	2	1	1	2
C. lari RM2100	1	2	1	1
C. upsaliensis RM3195	1	1	1	1
C. fetus fetus 82-40	3	2	1	0
C. fetus venerealis NCTC10354	3	2	1	0
C. sputorum INTA08/209	2	2	1	0
C. rectus RM3267	2	5	1	1
C. hominis ATCC-BAA 381	1	1	1	0
C. curvus 525.92	3	2	1	0
C. concisus 138.26	3	1	1	0
C. showae RM3277	2	2	1	0
C. gracilis RM3268	2	1	1	0
A. butzleri RM4018	2	1	1	0
A. nitrofigilis DSM7299	3	1	1	0
S. barnesii SES3	1	1	1	1
S. deleyianum DSM6946	1	1	1	1

4.7 References

[20] N. A. Moran, *Cell* **2002**, *108*, 583–586.

[130] I. Nachamkin, C. M. Szymanski, M. J. Blaser, et al., *Campylobacter*. ASM Press, **2008**.

[131] J. A. Wagenaar, M. A. van Bergen, M. J. Blaser, R. V. Tauxe, D. G. Newell, J. P. van Putten, *Clin. Infect. Dis.* **2014**, *58*, 1579–1586.

[132] C. T. Parker, W. G. Miller, S. T. Horn, A. J. Lastovica, *BMC microbiology* **2007**, *7*, 1.

[133] M Gürtler, T Alter, S Kasimir, K Fehlhaber, *Epidemiology and infection* **2005**, *133*, 1081–1087.

[134] S. K. Sheppard, X. Didelot, G. Meric, A. Torralbo, K. A. Jolley, D. J. Kelly, S. D. Bentley, M. C. Maiden, J. Parkhill, D. Falush, *Proceedings of the National Academy of Sciences* **2013**, *110*, 11923–11927.

[135] C. P. Skarp-de Haan, A. Culebro, T. Schott, J. Revez, E. K. Schweda, M.-L. Hänninen, M. Rossi, *BMC genomics* **2014**, *15*, 1.

[136] A. L. Jaime, S. Joan, B. Lee, S. Nancy, M. H. Sydney, L. Eleanor, R. Roshan, M. Laurene, *Clinical Infectious Diseases* **2002**, *34*, e59–e60.

[137] K Hayashi, A Tazumi, S Nakanishi, T Nakajima, K Matsubara, H Ueno, J. Moore, B. Millar, M Matsuda, *World Journal of Microbiology and Biotechnology* **2012**, *28*, 2403–2410.

[138] M Veron, R Chatelain, *International Journal of Systematic and Evolutionary Microbiology* **1973**, *23*, 122–134.

[139] G. Mshelia, J. Amin, Z Woldehiwet, R. Murray, G. Egwu, *Reproduction in Domestic Animals* **2010**, *45*, e221–e230.

[140] M. A. van Bergen, K. E. Dingle, M. C. Maiden, D. G. Newell, L. van der Graaf-Van Bloois, J. P. van Putten, J. A. Wagenaar, *J. Clin. Microbiol.* **2005**, *43*, 5888–5898.

[141] M. Skirrow, *Journal of comparative pathology* **1994**, *111*, 113–149.

[142] S. On, B Holmes, *Journal of clinical microbiology* **1992**, *30*, 746–749.

[143] W. Tee, M. Luppino, S. Rambaldo, *Clinical infectious diseases* **1998**, *27*, 1544–1545.

[144] S. On, H. Atabay, J. Corry, C. Harrington, P Vandamme, *International Journal of Systematic and Evolutionary Microbiology* **1998**, *48*, 195–206.

[145] S. M. Man, *Nat Rev Gastroenterol Hepatol* **2011**, *8*, 669–685.

[146] J. Siqueira, I. Rocas, *International endodontic journal* **2003**, *36*, 174–180.

[147] D. Linton, R. J. Owen, J. Stanley, *Res. Microbiol.* **1996**, *147*, 707–718.

[148] J. Butler, I. MacCallum, M. Kleber, I. A. Shlyakhter, M. K. Belmonte, E. S. Lander, C. Nusbaum, D. B. Jaffe, *Genome research* **2008**, *18*, 810–820.

[149] D. R. Zerbino, E. Birney, *Genome Res.* **2008**, *18*, 821–829.

[150] M. T. Swain, I. J. Tsai, S. A. Assefa, C. Newbold, M. Berriman, T. D. Otto, *Nat Protoc* **2012**, *7*, 1260–1284.

[151] S. C. Clark, R. Egan, P. I. Frazier, Z. Wang, *Bioinformatics* **2013**, bts723.

[152] R. K. Aziz, D. Bartels, A. A. Best, M. DeJongh, T. Disz, R. A. Edwards, K. Formsma, S. Gerdes, E. M. Glass, M. Kubal, F. Meyer, G. J. Olsen, R. Olson, A. L. Osterman, R. A. Overbeek, L. K. McNeil, D. Paarmann, T. Paczian, B. Parrello, G. D. Pusch, C. Reich, R. Stevens, O. Vassieva, V. Vonstein, A. Wilke, O. Zagnitko, *BMC Genomics* **2008**, *9*, 75.

[153] L. Li, C. J. Stoeckert, D. S. Roos, *Genome research* **2003**, *13*, 2178–2189.

[154] L. Chen, J. Yang, J. Yu, Z. Yao, L. Sun, Y. Shen, Q. Jin, *Nucleic acids research* **2005**, *33*, D325–D328.

[155] A. Conesa, S. Götz, J. M. García-Gómez, J. Terol, M. Talón, M. Robles, *Bioinformatics* **2005**, *21*, 3674–3676.

[156] R. C. Edgar, *Nucleic acids research* **2004**, *32*, 1792–1797.

[157] S.-R. Jun, G. E. Sims, G. A. Wu, S.-H. Kim, *Proceedings of the National Academy of Sciences* **2010**, *107*, 133–138.

[158] E. Pruesse, C. Quast, K. Knittel, B. M. Fuchs, W. Ludwig, J. Peplies, F. O. Glöckner, *Nucleic acids research* **2007**, *35*, 7188–7196.

[159] E. Paradis, J. Claude, K. Strimmer, *Bioinformatics* **2004**, *20*, 289–290.

[160] B. Murrell, J. O. Wertheim, S. Moola, T. Weighill, K. Scheffler, S. L. K. Pond, *PLoS Genet* **2012**, *8*, e1002764.

[161] T. N. Petersen, S. Brunak, G. von Heijne, H. Nielsen, *Nature methods* **2011**, *8*, 785–786.

[162] J. D. Bendtsen, L. Kiemer, A. Fausbøll, S. Brunak, *BMC microbiology* **2005**, *5*, 58.

[163] D. Charif, J. R. Lobry in *Structural approaches to sequence evolution*, Springer, **2007**, pp. 207–232.

[164] O. Nenadic, M. Greenacre, **2007**.

[165] E. P. Rocha, A. Danchin, *TRENDS in Genetics* **2002**, *18*, 291–294.

[166] N. A. Moran, G. R. Plague, *Current opinion in genetics & development* **2004**, *14*, 627–633.

[167] D. Medini, C. Donati, H. Tettelin, V. Masignani, R. Rappuoli, *Current opinion in genetics & development* **2005**, *15*, 589–594.

[168] M. D. Collins, M. R. Jovita, P. A. Lawson, E. Falsen, G. Foster, *International Journal of Systematic and Evolutionary Microbiology* **1999**, *49*, 1381–1385.

[169] M. Asakura, W. Samosornsuk, M. Taguchi, K. Kobayashi, N. Misawa, M. Kusumoto, K. Nishimura, A. Matsuhisa, S. Yamasaki, *Microbial pathogenesis* **2007**, *42*, 174–183.

[170] K. A. Talukder, M. Aslam, Z. Islam, I. J. Azmi, D. K. Dutta, S. Hossain, A. Nur-E-Kamal, G. B. Nair, A. Cravioto, D. A. Sack, et al., *Journal of clinical microbiology* **2008**, *46*, 1485–1488.

[171] G. Ripabelli, M. Tamburro, F. Minelli, A. Leone, M. L. Sammarco, *Comparative immunology microbiology and infectious diseases* **2010**, *33*, 355–364.

[172] J. d. S. Quetz, I. F. Lima, A. Havt, M. M. Prata, P. A. Cavalcante, P. H. Medeiros, D. A. Cid, M. L. Moraes, L. C. Rey, A. M. Soares, et al., *Journal of medical microbiology* **2012**, *61*, 507–513.

[173] D Purdy, C. Buswell, A. Hodgson, K McAlpine, I Henderson, S. Leach, *Journal of medical microbiology* **2000**, *49*, 473–479.

[174] D. Jain, K. N. Prasad, S. Sinha, N. Husain, *Journal of medical microbiology* **2008**, *57*, 267–272.

[175] P. M. Moolhuijzen, A. E. Lew-Tabor, B. M. Wlodek, F. G. Agüero, D. J. Comerci, R. A. Ugalde, D. O. Sanchez, R. Appels, M. Bellgard, *BMC microbiology* **2009**, *9*, 86.

[176] S. Kienesberger, H. Sprenger, S. Wolfgruber, B. Halwachs, G. G. Thallinger, G. I. Perez-Perez, M. J. Blaser, E. L. Zechner, G. Gorkiewicz, *PloS one* **2014**, *9*, e85491.

[177] A. J. Lawson, D. Linton, J. Stanley, *Microbiology* **1998**, *144*, 2063–2071.

[178] T. M. Wassenaar, W. Gaastra, *FEMS microbiology letters* **2001**, *201*, 1–7.

[179] D. A. Alfredson, V. Korolik, *FEMS Microbiology Letters* **2007**, *277*, 123–132.

[180] M. E. Konkel, L. A. Joens, *Infection and immunity* **1989**, *57*, 2984–2990.

[181] A. Mooney, C. Byrne, M. Clyne, K. Johnson-Henry, P. Sherman, B. Bourke, *Cellular microbiology* **2003**, *5*, 835–847.

[182] D. McDougald, L. Gong, S. Srinivasan, E. Hild, L. Thompson, K. Takayama, S. A. Rice, S Kjelleberg, *Antonie Van Leeuwenhoek* **2002**, *81*, 3–13.

[183] E. C. Hagan, H. L. Mobley, *Infection and immunity* **2007**, *75*, 3941–3949.

[184] C. J. Yeoman, S. Yildirim, S. M. Thomas, A. S. Durkin, M. Torralba, G. Sutton, C. J. Buhay, Y. Ding, S. P. Dugan-Rocha, D. M. Muzny, et al., *PLoS One* **2010**, *5*, e12411.

[185] C. A. Kumamoto, *Current opinion in microbiology* **2008**, *11*, 325–330.

[186] R. P. Darveau, *Nature Reviews Microbiology* **2010**, *8*, 481–490.

[187] M. Vianna, S Holtgraewe, I Seyfarth, G Conrads, H. Horz, *Journal of bacteriology* **2008**, *190*, 3779–3785.

[188] D. Kara, S. B. Luppens, J. M. Cate, *European journal of oral sciences* **2006**, *114*, 58–63.

[189] J. S. McLean, S. J. Fansler, P. D. Majors, K. McAteer, L. Z. Allen, M. E. Shirtliff, R. Lux, W. Shi, *PloS one* **2012**, *7*, e32219.

[190] M. Desvaux, M. Hébraud, R. Talon, I. R. Henderson, *Trends in microbiology* **2009**, *17*, 139–145.

[191] C. Song, A. Kumar, M. Saleh, *Genomics proteomics & bioinformatics* **2009**, *7*, 37–46.

[192] I. J. Schalk, M. Hannauer, A. Braud, *Environmental microbiology* **2011**, *13*, 2844–2854.

[193] N. Noinaj, M. Guillier, T. J. Barnard, S. K. Buchanan, *Annual review of microbiology* **2010**, *64*, 43.

[194] P. M. Sharp, E. Bailes, R. J. Grocock, J. F. Peden, R. E. Sockett, *Nucleic acids research* **2005**, *33*, 1141–1153.

[195] I. Bahir, M. Fromer, Y. Prat, M. Linial, *Molecular systems biology* **2009**, *5*.

[196] B. Heras, S. R. Shouldice, M. Totsika, M. J. Scanlon, M. A. Schembri, J. L. Martin, *Nature Reviews Microbiology* **2009**, *7*, 215–225.

[197] L. L. Graham, *Canadian journal of microbiology* **2002**, *48*, 995–1007.

[198] S. M. Man, L. Zhang, A. S. Day, S. T. Leach, D. A. Lemberg, H. Mitchell, *Inflammatory bowel diseases* **2010**, *16*, 1008–1016.

[199] T. Lefébure, M. J. Stanhope, *Genome research* **2009**, *19*, 1224–1232.

[200] M. J. Blaser, Z. Pei, *Journal of Infectious Diseases* **1993**, *167*, 372–377.

[201] K. C. Ray, Z.-C. Tu, R. Grogono-Thomas, D. G. Newell, S. A. Thompson, M. J. Blaser, *Infection and immunity* **2000**, *68*, 5663–5667.

Conclusión del Capítulo 4

El género *Campylobacter* esta compuesto por un gran número de especies que presentan un amplio rango de hospederos y tropismo por distintos tejidos, así como un grado de patogenicidad diferencial para el humano.

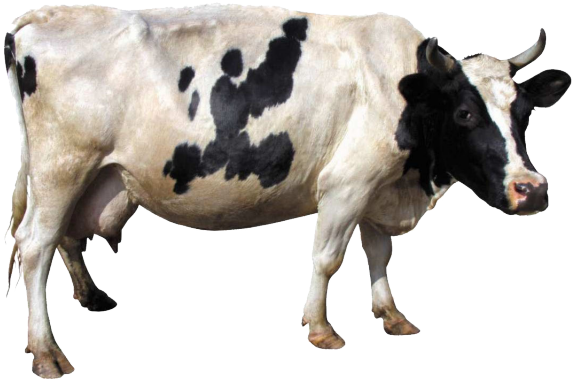
En este trabajo se realizó una comparación de genomas representativos de todas las especies secuenciadas hasta ese momento, con el objetivo de identificar características genómicas relacionadas a la adaptación de las especies a distintos nichos y la evolución de su patogenicidad.

Como contribución principal identificamos dos grandes grupos de especies distinguidas por sus repertorios de genes de virulencia, los cuales presentan claras señales de transferencia horizontal entre alguno de sus miembros. En general, nuestros resultados evidenciaron que la emergencia de la patogenicidad esta correlacionada con la adquisición de genes de virulencia a través de transferencia horizontal desde otras bacterias y entre los miembros de *Campylobacter*. Además, la preferencia de nicho puede ser explicada por evolución no sinónima de un conjunto de genes que codifican proteínas como DSBs y CiaB, y por diferencias en el contenido en GC y el tamaño de los genomas y el secretoma.

En este trabajo se abordó por primera vez una comparación genómica utilizando el total de la diversidad taxonómica conocida del género *Campylobacter* y a su vez se presentó el primer genoma de la especie *C. sputorum*. Se puntualiza además la necesidad de secuenciar genomas de especies no modelo para proporcionar un panorama no sesgado de la diversidad geómica de las bacterias, que puede ser utilizada para obtener información valiosa acerca de sus mecanismos evolutivos.

5

Genomes uncover cattle-to-human transmission of *Campylobacter fetus* in Uruguay



Citation:

Iraola G*, Betancor L, Calleros L,Gadea P, Algorta G, Galeano S, Muxi P, Greif G, Pérez R. (2015) **A rural worker infected with a bovine-prevalent genotype of *Campylobacter fetus* subsp. *fetus* supports zoonotic transmission and inconsistency of MLST and whole-genome typing.** *European Journal of Clinical Microbiology and Infectious Disease*. 4(8):1593-6.

* Corresponding author

5.1 Abstract

Whole-genome characterization in clinical microbiology enables to detect trends in infection dynamics and disease transmission. Here, we report a case of bacteraemia due to *Campylobacter fetus* subsp. *fetus* in a rural worker under cancer treatment that was diagnosed with cellulitis; the patient was treated with antibiotics and recovered. The routine typing methods were not able to identify the microorganism causing the infection, so it was further analysed by molecular methods and whole-genome sequencing. The multi-locus sequence typing (MLST) revealed the presence of the bovine associated ST-4 genotype. Whole-genome comparisons with other *C. fetus* strains revealed an inconsistent phylogenetic position based on the core genome, discordant with previous ST-4 strains. To the best of our knowledge, this is the first *C. fetus* subsp. *fetus* carrying the ST-4 isolated from humans and represents a probable case of zoonotic transmission from cattle.

5.2 Case presentation

In October 2010, a 64-year-old male rural worker with treated hypothyroidism and vascular complications was diagnosed with a low-risk non-Hodgkin mantle cell lymphoma, presenting polyps in the colon and the small intestine. Computerised axial tomography revealed no adenopathies and bone marrow biopsy showed no infiltrations. The patient resulted CD5+, CD23+, border-line for D1 cyclin, had low expression of Ki67 and was positive for t(11;14). Complete remission was obtained after six cycles of R-CHOP chemotherapy. In a posterior routine control (November 2012), fibre gastroscopy and colonoscopy revealed various lesions, confirming the first relapse for the mantle cell lymphoma. A bendamustineÐrituximab protocol was used for treatment re-induction. After finishing the treatment (April 2013), no adenopathies were evident and bone marrow myelography and immunophenotyping revealed no infiltrations. Haemogram, hepatic and renal functions were normal. The patient was admitted for autologous transplantation of haematopoietic progenitors under partial remission. In July 2013, cellulitis was observed in the right leg. Two independent

blood culture sets were performed using the BacT/ALERT 3D system (3D) (bioMérieux, Inc., Durham, NC) and, 2 days after incubation, bacterial growth was detected. Isolation was done in tryptic soy agar (Oxoid, Hampshire, England), supplemented with 5% of sheep blood incubated at 37°C. After 72 h of incubation, small, punctiform and brilliant colonies were observed on plates incubated under aerobic conditions. The Gram stain of the colonies showed spiral Gram-negative rods, and the catalase and oxidase tests were positive. As these results were consistent with the presence of *Campylobacter* sp., the strain was isolated on blood agar plates under microaerobic atmosphere. Grey, flat, spreading irregular colonies were obtained after 48 h of culture. The VITEK 2 system (bioMérieux, Marcy L'Etoile, France) with the NH card identified the strain as *Campylobacter fetus/coli* but was unable to distinguish between both species. For species identification, a fragment of the 16S rRNA gene was amplified and sequenced using previously described conditions [147]. Comparative sequence analysis unequivocally identified the isolate as *C. fetus*. The patient was treated with metronidazole and levofloxacin, and had a good evolution. The isolated strain was named H1-UY for subsequent characterization.

5.3 Molecular and genomic characterization

The species *C. fetus* is assumed to be divided into the subspecies *C. fetus* subsp. *venerealis* (Cfv), *C. fetus* subsp. *fetus* (Cff) and *C. fetus* subsp. *testudinum* (Cft). The analysis of the 16S rRNA gene has been widely applied and validated to identify *Campylobacter* strains at the species level; however, this genetic marker is too conserved to differentiate Cff from Cfv. Accordingly, the partial 16S gene sequence from H1-UY presented 100% identity with Cff and Cfv strains available in sequence databases. The subspecies Cft was described as genetically divergent from Cfv and Cff [202], so the ten nucleotide differences found in the partial 16S of H1-UY with respect to the reference Cft 02-427 was the first evidence that H1-UY should be classified as Cff or Cfv but not Cft. For additional characterization of H1-UY, we used three published polymerase chain reaction (PCR) assays that target different markers useful to distinguish between Cff and Cfv [86, 203, 204]. These assays consistently identified H1-UY as Cff.

To proceed further in the molecular characterization of H1-UY and procure additional evidence for subspecies determination, the whole genome was sequenced on an Illumina MiSeq platform, generating 1,394,690 paired-end reads (2? 150 cycles) after quality filtering. The resulting library was assembled with Velvet [149] and improved with PAGIT [150], producing 34 contigs with an average coverage of 137-fold. The resulting high-quality draft genome was annotated with RAST [152] and deposited in the GenBank under accession number JYCP00000000. The best Blast reciprocal hit approach was used to identify the core genome (nucleotide identity >50% and query coverage >80%) between H1-UY and previously available genomes for Cff, Cfv and Cft strains. The phylogenetic reconstruction using 25,166 polymorphic sites in the core genome revealed that the human isolate Cff 82-40 was the closest strain to H1-UY and that Cft deeply branched from all Cff and Cfv strains (Fig. 5.1). This result confirmed that H1-UY can be classified as Cff. The standard multi-locus sequence typing (MLST) scheme for *C. fetus* was applied to H1-UY, which was assigned to ST-4, so far a non-reported genotype among human-derived Cff strains. The presence of ST-4 was also confirmed by retrieving and analyzing the MLST genes from the H1-UY genome.

5.4 Discussion

The species *C. fetus* is a renowned pathogen worldwide that produces considerable economic losses, mainly in bovine and ovine productive chains for being a primary cause of ruminant infertility and abortions. The increasing bacterial typing effort in clinical microbiology and the development of more powerful molecular techniques have revealed a previously underrated role of *C. fetus* in human infections [131]. Among *C. fetus* subspecies, Cff presents the greatest incidence in human infections and shows a wide host range, including sheep, cattle and reptiles. Cfv is host-restricted, being isolated almost exclusively from the bovine genital tract and causing fertility problems. Cft has been recently proposed based on the characterization of genetically divergent strains isolated from reptiles and ill humans [202].

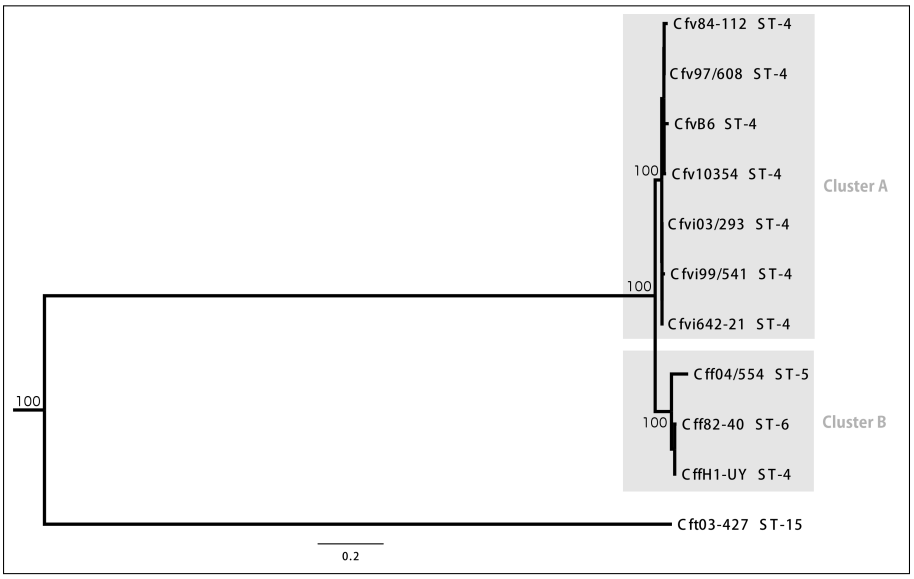


Figure 5.1: Maximum likelihood tree inferred from core genomes. The grey squares highlight genomes belonging to cluster "A" and cluster "B". Bootstrap values are presented for most significant nodes. Tree branches lengths are expressed in substitutions per site.

Since the standardization of the MLST scheme for *C. fetus*, it has been widely applied to genetically characterize strains isolated from different hosts and, in particular, the genotype ST-4 has been frequently detected in Cfv and to a lesser extent in Cff isolated from cattle, which leads to propose ST-4 as a bovine-associated genotype [140]. This genotype has also been detected among a very small number of Cfv strains isolated from human vaginal discharges but, to date, no human derived Cff strain has been identified as ST-4 (Fig. 5.2). The availability of whole-genome sequences also allowed to compare phylogenies based on MLST and core genomes. In a recent study, the comparison of Cff and Cfv core genomes evidenced the presence of two genomic clusters that correlate with MLST typing [205]. Cluster "B" is comprised of bovine derived strains carrying the ST-4 and exclusively assigned as Cfv or Cfvi by amplified fragment length polymorphism (AFLP),while the strains carrying other STs comprised cluster "B"; the strain Cff 82-40 that has the ST-6 was placed in cluster "B". We replicated this analysis and found that H1-UY should also be placed in cluster "B" based

on its genomic relatedness with Cff 82-40, in spite of being ST-4. Beyond reporting the first human-derived Cff strain with the ST-4, this work evidenced that MLST and core genome phylogenies are not always consistent, unlike previously reported results [205]. The ST-4 and ST-6 are defined just by a synonymous single nucleotide transition (G708A) in the *uncA* gene and the analysis of its genomic context in H1-UY showed total synteny conservation with Cff 82-40, suggesting that ST-4 arose by point mutation in this case. Despite that MLST and core genome tree topologies are highly correlated, the low genetic distance between some MLST genotypes, like ST-6 and ST-4, prevents to ensure that MLST totally reflects the core genome. Our results underscore the importance of complementing MLST with whole-genome sequencing for typing *C. fetus* subspecies from human isolates, which seems crucial to elucidate the epidemiology of *C. fetus* and its role as a zoonotic pathogen. Furthermore, since the completion of the Cff 82-40 genome in 2006, this is the first published Cff genome isolated from humans, which contributes to the generation of genomic data for upcoming genetics studies.

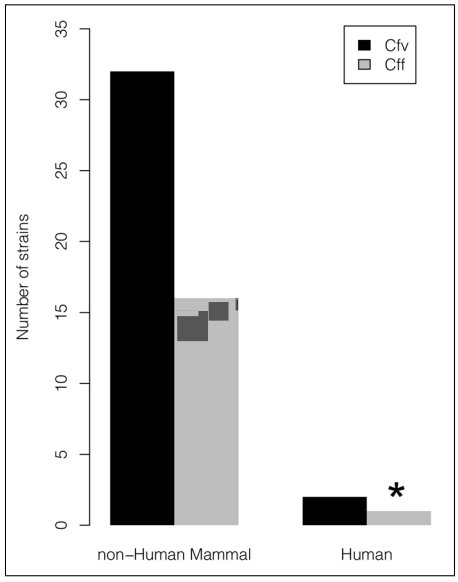


Figure 5.2: Bibliographic revision of *Campylobacter fetus* strains genotyped as ST-4 considering host and subspecies. The asterisk indicates that H1-UY is the first *C. fetus* subsp. *fetus* (Cff) being ST-4 isolated from humans.

In spite of being rare, the association between cellulitis and *C. fetus* bacteraemia has been well documented [206]. In the case described here, cellulitis was diagnosed 3 months after the patient finished chemotherapy against a relapsing mantle cell lymphoma, in accordance with previous studies which demonstrated that most *C. fetus* infections have other underlying disease, like diabetes, human immunodeficiency virus (HIV), cancer or another risk factor associated with immunosuppression [131, 140, 205–207]. In addition, the patient was exposed to occupational risk factors because his activities were related to rural work, in daily contact with cattle. Despite the lack of information about the differential incidence of *C. fetus* in rural and urban settings, the transmission from animals to humans via direct contact or contaminated products has been proposed, and most human infections seem to have a bovine origin [131], so this route of transmission is the most probable in this case.

5.5 References

[86] G. Iraola, M. Hernandez, L. Calleros, F. Paolicchi, S. Silveyra, A. Velilla, L. Carretto, E. Rodriguez, R. Perez, *J. Vet. Sci.* **2012**, *13*, 371–376.

[131] J. A. Wagenaar, M. A. van Bergen, M. J. Blaser, R. V. Tauxe, D. G. Newell, J. P. van Putten, *Clin. Infect. Dis.* **2014**, *58*, 1579–1586.

[140] M. A. van Bergen, K. E. Dingle, M. C. Maiden, D. G. Newell, L. van der Graaf-Van Bloois, J. P. van Putten, J. A. Wagenaar, *J. Clin. Microbiol.* **2005**, *43*, 5888–5898.

[147] D. Linton, R. J. Owen, J. Stanley, *Res. Microbiol.* **1996**, *147*, 707–718.

[149] D. R. Zerbino, E. Birney, *Genome Res.* **2008**, *18*, 821–829.

[150] M. T. Swain, I. J. Tsai, S. A. Assefa, C. Newbold, M. Berriman, T. D. Otto, *Nat Protoc* **2012**, *7*, 1260–1284.

[152] R. K. Aziz, D. Bartels, A. A. Best, M. DeJongh, T. Disz, R. A. Edwards, K. Formsma, S. Gerdes, E. M. Glass, M. Kubal, F. Meyer, G. J. Olsen, R. Olson, A. L. Osterman, R. A. Overbeek, L. K. McNeil, D. Paarmann, T. Paczian, B. Parrello, G. D. Pusch, C. Reich, R. Stevens, O. Vassieva, V. Vonstein, A. Wilke, O. Zagnitko, *BMC Genomics* **2008**, *9*, 75.

[202] C. Fitzgerald, Z. C. Tu, M. Patrick, T. Stiles, A. J. Lawson, M. Santovenia, M. J. Gilbert, M. van Bergen, K. Joyce, J. Pruckler, S. Stroika, B. Duim, W. G. Miller, V. Loparev, J. C. Sinnige, P. I. Fields, R. V. Tauxe, M. J. Blaser, J. A. Wagenaar, *Int. J. Syst. Evol. Microbiol.* **2014**, *64*, 2944–2948.

[203] S. Hum, K. Quinn, J. Brunner, S. L. On, *Aust. Vet. J.* **1997**, *75*, 827–831.

[204] C. Abril, E. M. Vilei, I. Brodard, A. Burnens, J. Frey, R. Miserez, *Clin. Microbiol. Infect.* **2007**, *13*, 993–1000.

[205] L. van der Graaf-van Bloois, W. G. Miller, E. Yee, M. Rijnsburger, J. A. Wagenaar, B. Duim, *J. Clin. Microbiol.* **2014**, *52*, 4183–4188.

[206] S. Ichiyama, S. Hirai, T. Minami, Y. Nishiyama, S. Shimizu, K. Shimokata, M. Ohta, *Clin. Infect. Dis.* **1998**, *27*, 252–255.

[207] L. Gazaigne, P. Legrand, B. Renaud, B. Bourra, E. Taillandier, C. Brun-Buisson, P. Lesprit, *Eur. J. Clin. Microbiol. Infect. Dis.* **2008**, *27*, 185–189.

Conclusión del Capítulo 5

En este trabajo se presenta el primer caso clínico reportado de *Campylobacter fetus* en Uruguay, causante de una bacteremia en un paciente inmunodeprimido por quimioterapia. La presentación del caso se complementó con la secuenciación del genoma completo de la cepa aislada, cuyo análisis permitió inferir un posible evento de transmisión zoonótica. También se determinó una incongruencia entre la tipificación por MLST y la basada en el genoma completo, planteando interrogantes acerca de la utilización del MLST como marcador asociado al hospedero.

Luego de reportado este primer caso, se han confirmado en distintos centros de salud nacionales siete casos más en dos años, un nuḿuero considerable dada la baja frecuencia de esta especie en humanos. Posiblemente el aumento de los casos no se deba a un fenómeno epidemiológico concreto sino a la optimización de los métodos de búsqueda y aislamiento a partir de la aparición del primer caso.

Actualmente, y mediante una colaboración con el Wellcome Trust Sanger Institute, hemos generado cerca de 200 genomas de *C. fetus* provenientes de cepas aisladas en distintos países y de distintos hospederos. El objetivo de este trabajo en curso es determinar qué variaciones en el genoma de esta especie han sido responsables de su diversificación y adaptación a distintos hospederos, y en particular elucidar los mecanismos de transmisión entre humanos y animales de producción como ovinos y bovinos. La reconsturcción de la historia poblacional de esta especie perimtirá también estimar los tiempos de diversificación e inferir el número y direccionalidad de los cambios de hospedero.

El análisis genómico a nivel poblacional permitirá elucidar las fuerzas evolutivas que modelan los genomas de esta especie, lo que permitirá un mejor conocimiento de su epidemiologŠaía. Este conocimiento podrá ser utilizado para generar mejores planes de diganóstico, control y tratamiento.

6

The sprinter genomes of *Campylobacter hyointestinalis*: yet another emerging pathogen

Citation:

Iraola G*, Levésque S, Kumar N, Naya H, Lawley TD*. (2016) **Rapidly evolving *Campylobacter hyointestinalis* subsp. *hyointestinalis* co-occurring in neighboring cattle farms.** *Microbial Genomics*. Under review.

* Corresponding authors

6.1 Abstract

Non-classical *Campylobacter* species are increasing viewed as emerging pathogens for humans and animals, although the biological reasons remain poorly understood. In this work we whole genome sequenced and performed a comprehensive analysis of 13 *C. hyointestinalis* strains isolated from healthy cattle (n=12) or a natural watercourse (n=1) on neighboring farms. Despite being geographically restricted, the *C. hyointestinalis* population displayed tremendous genomic diversity. Genome-wide recombination rates in *C. hyointestinalis* were significantly higher than in its sister species *C. fetus*, suggesting that recombination is a major force shaping *C. hyointestinalis* genome diversity. In particular, recombinant regions harbored genes of basal metabolic pathways such as energy production. We also observed an extremely high substitution rate ($1.4x10^{-3}$ s/s/y) highlighting a second major force in driving *C. hyointestinalis* evolution. Whole genome phylogenetic analysis identified three evolutionary lineages each with distinct evolutionary patterns and defined by unique patterns of gene gain/loss such as those functioning in LPS biosynthesis. We also found that distinct phylogenetic lineages co-occurred in the same farm implying frequent transmission between farms and environmental sources. Based on our analysis, we propose that high genomic plasticity supports the adaptive potential of *C. hyointestinalis* metabolism and host interactions for its dual role as a commensal in cattle and emerging pathogen in humans.

6.2 Introduction

The genus *Campylobacter* consists of a diverse group of bacteria currently classified into 25 species and 12 subspecies. Among them, *C. jejuni* and *C. coli* have drawn the most attention because they are the leading causes of human gastroenteritis worldwide [208]. However, the recent development of sensitive molecular diagnostic methods and an increased clinical awareness of campylobacteriosis have highlighted other neglected *Campylobacter* species as causative agents of human and animal infections [145]. In particular, *C. hyointestinalis* was first

isolated from swine with proliferative enteritis [209] and has since been found associated with infections in humans and a wide variety of wild, farm and domestic mammals including cattle, pigs, dogs, hamsters, deer, reindeer and sheep [145]. Interestingly, *C. hyointestinalis* species has been found in both healthy and diseased hosts. These observations raise the possibility that *C. hyointestinalis* is an emerging zoonotic pathogen that can cause opportunistic infections in humans [210, 211].

C. hyointestinalis is divided in two subspecies, namely *C.hyointestinalis* subsp. *lawsonii* and *C. hyointestinalis* subsp. *hyointestinali*s, based on genetic and phenotypic traits [212, 213]. While *C. hyointestinalis* subsp. *hyointestinalis* has a broad host range, *C. hyointestinalis* subsp. *lawsonii* is restricted to pigs. Genetic and protein analysis have suggested that *C. hyointestinalis* harbours considerable intra-species genetic diversity [214] which could facilitate its adaptation to diverse hosts and environments [212, 215]. However, there remains a lack of genomic data for *C. hyointestinalis* so the phylogenetic diversity and genetic mechanisms underlying any potential variability have not been explored at whole-genome level.

In this work we produced the first whole-genome sequences for *Campylobacter hyointestinalis* subsp. *hyointestinalis* strains that were isolated from healthy cattle and an environmental watercourse sampled on dairy and beef farms located around Sherbrooke, Québec, Canada. Through comparative genomics and phylogenetics we highlight that both recombination and point mutation are significant forces shaping the evolution and transmission of commensal *C. hyointestinalis*.

6.3 Methods

6.3.1 Sampling and bacterial isolation. Samples were collected as described previously. Briefly, cattle feces samples were transported in Enteric Plus medium (Meridian Bioscience Inc, Ohio, USA) and processed on the same day. About 1-2 g of each fecal sample were transferred to 25 ml of Preston selective enrichment broth (Oxoid, Nepean, Ontario, Canada) and incubated 3-4 h at 37°C and then transferred to 42°C and incubated for 48 h. After incubation, 20 ul were streaked on a Karmali plate (Oxoid) and incubated at 42°C for 48 h. For environ-

mental water, 3000 ml of water were collected and transported on ice to the laboratory, held at 4°C and tested within 24 h. Water was filtered through a 0.45 um pore-size membrane filter and Preston broth and Karmali plate were used as above to isolate *Campylobacter*.

6.3.2 Whole genome sequencing. Cells were pelleted from culture plates and phosphate-buffered saline (PBS). Genomic DNA preparation was performed using a BioRobot M48 (Qiagen). DNA was prepared and sequenced using the Illumina Hi-Seq platform with library fragment sizes of 200-300 bp and a read length of 100 bp at the Wellcome Trust Sanger Institute, as previously described [216]. Each sequenced genome was de novo assembled with Velvet [149], SSPACE v. 2.0 [217] and GapFiller v 1.1 [218] using an in-house pipeline developed at the Wellcome Trust Sanger Institute. Resulting contigs were annotated using Prokka [219].

6.3.3 Genome diversity analyses. First, a Maximum Likelihood phylogeny was built with MEGA6 [220] from the concatenated alignment of 40 universal proteins for prokaryotes, retrieved with FetchMG [221] from the genomes sequenced in this study and *C. hyointestinalis*, *C. fetus* and *C. iguaniorum* genomes from public databases (Table S3). This confirmed that the genomes sequenced in this study belong to the species *C. hyointestinalis*. Average Nucleotide Identity (ANI) was calculated using BLAST [79] by implementing an in-house R function based on the algorithm detailed in Konstantinidis and Tiedje [222].

Second, all known alleles for the seven housekeeping genes used in *C. hyointestinalis* MLST scheme [213] (*aspA*, *atpA*, *glnA*, *gltA*, *glyA*, *pgm* and *tkt*) were downloaded from the PubMLST database (http://pubmlst.org). Each genome was screened using in-house Perl scripts to recover MLST genes and compare them with published alleles to determine the presence of novel alleles and reconstruct sequence types (STs). A Neighbor-Joining phylogeny was built in MEGA6 [220] by concatenating the seven genes. Recombination along MLST genes was assessed with Gubbins [223].

To compare the gene conservation among *C. hyointestinalis* genomes sequenced in this study and the public genome of DSM 19053, a BLAST [79] database was created with all genes annotated in DSM 19053 as

reference and the remaining genomes were queried using BLASTp. A certain gene was flagged as present if shared >50% of identity and >80% of alignment length. Results were visualized with Circos [224] using DSM 19053 as reference.

6.3.4 Whole-genome phylogeny and population structuring. The core and accessory genomes of *C. hyointestinalis* were estimated using Roary [225] at 90% identity and 99% coverage. The concatenated core genes were aligned with PRANK [226] and Gubbins was used to remove recombinant blocks. The clonal frame alignment was used to build a maximum likelihood tree using RAxML [227] with the GTR model and 1000 bootstrap replicates. The resulting tree was used to perform a population structure analysis with BAPS using default parameters [228]. The Jaccard index using patterns of accessory genes presence/absence was calculated with ade4 package in R.

6.3.5 Recombination and substitution rates. Recombination and substitution rates. For assessing the effect of recombination over the genomes of *C. hyointestinalis*, a comparison with *C. fetus* genomes was performed due to both species share an immediate common ancestor in the *Campylobacter* phylogeny and also *C. fetus* possesses similar reservoirs in nature than *C. hyointestinalis*. Thirteen *C. fetus* genomes were obtained from public databases. For obtaining the accessory and core genomes for the joint dataset of *C. hyointestianlis* and *C. fetus* the annotated genes were inputed to Roary [225] to identify the accessory and core genomes at 70% of identity and 99% of presence coverage. The concatenated core genes were aligned with PRANK [226] and Gubbins was used to infer recombinant blocks with default parameters [229]. The sequences in recombinant blocks were extracted using in-house R scripts and the contained genes were functionally annotated and assigned to metabolic pathways with KAAS [230]. The mutation rate for C. hyointestinalis and C. fetus was estimated separately using BEAST2 [231]. The GTR model was used assuming a constant population prior and a relaxed molecular clock. In order to ensure proper convergence, four independent Markov chain Monte Carlo (MCMC) chains were run, each of 50,000,000 states. From these, an initial 10% (5,000,000 states)

was removed as a burn-in and then chains were joined using LogCombiner by taking a sample every 1,000 states.

6.4 Results

6.4.1 Genomic diversity. We whole genome sequenced, assembled and annotated 13 *Campylobacter* strains isolated between July 2005 and November 2007 from cattle on 12 different farms and one environmental watercourse, from a geographically restricted area defined by a radius of 150 km around the city of Sherbrooke, Quèbec, Canada (Table 1, Fig. 6.1A). The strains were identified as *C. hyointestinalis* by 16S rRNA gene identity and phylogenetic analyses using 40 universal bacterial proteins [221] (Supp. Fig. 6.1A). Also, the Average Nucleotide Identity (ANI) values were >95% indicating that sequenced genomes belong to the same species, except for *C. hyointestinalis* DSM 19053 (public genome) which showed ANI values below 95% (Fig. 6.1B); DSM 19053 was originally isolated from a diseased pig so it is considered a pathogenic strain. The genomes of *C. hyointestinalis* varied from 1.73 to 2.0 Mb in length and from 32.5 to 35.9 in average GC content (Table S1). Based on the phylogeny built by concatenating MLST genes extracted from the PubMLST database (`http://pubmlst.org/campylobacter/`), the 13 sequenced strains were unequivocally subtyped as *C. hyointestinalis* subsp. *hyointestinalis* (Fig. 6.1B). Currently, 9 different STs have been described for *C. hyointestinalis* subsp. *lawsonii* and 122 for *C. hyointestinalis* subsp. *hyointestinalis*. Each sequenced genome represents a distinct ST and 7 of them are novel STs (Table 1). The genome of the reference strain DSM 19053 belongs to ST-27. The strain 006A-0180 presented two novel alleles for genes *glnA* and *pgm*. Strains 006A-0113 and 006A-0091 presented two novel alleles for gene *aspA*, while 006A-0191 and 006A-0059 had novel sequences for gene *tkt*. Two strains (006A-0178 and 006A-0063) represented new STs by novel combinations of previously described alleles. The recombination analysis performed over MLST genes revealed strain-specific recombination signals in *tkt* and *pgm* genes of strain 006A-0180.

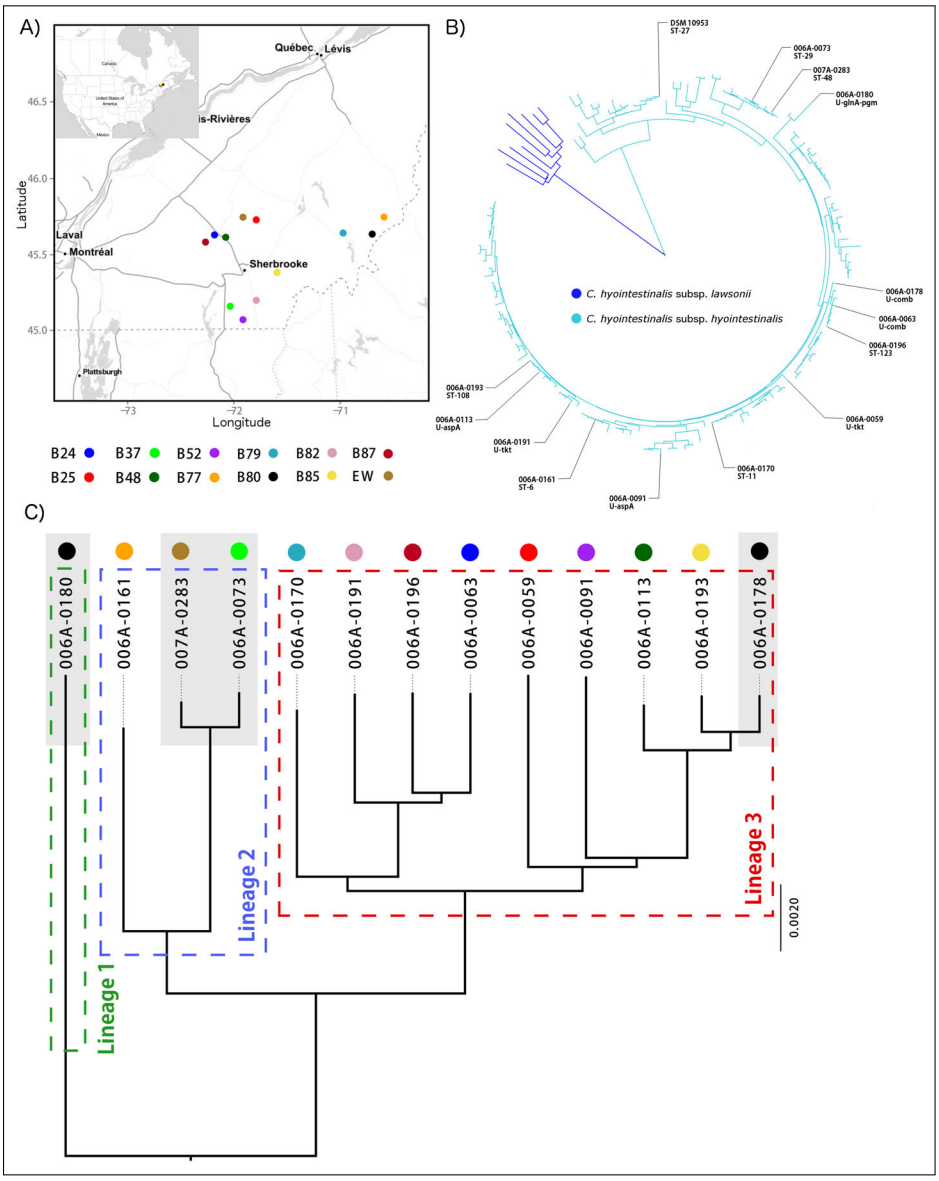


Figure 6.1: Geographic distribution, structuring and transmission. A) Map showing the geographic points were samples were taken. B) Phylogenetic tree using concatenated MLST alleles available in PubMLST database. Blue branches represent *C. hyointestinalis* subsp. *lawsonii* and light blue branches represent *C. hyointestinalis* subsp. *hyointestinalis*. Strains sequenced in this study are labeled in black as well as the type strain DSM 19053. When new STs were detected they are labeled with "U" (unknown) and the names of genes presenting novel alleles. When the new ST was created by a novel combination of previously reported alleles they were tagged as "U-comb". C) Phylogenetic tree based on non-recombined regions (clonal frame) annotated with geographic locations and highlighting the genomic lineages identified through BAPS analysis. Strains highlighted in grey were isolated at the same location or evidence the transmission between farms and the environment.

Each of the 13 genomes harboured most of the virulence-associated genes present in the pathogenic DSM 19053, including a conserved CiaB invasin gene, a type IV pilus system, cytolethal distending toxin subunits and several clusters coding for flagellar genes spread in the genome. All strains contained CRISPR loci including variable-length direct repeats and spacers; conserved CRISPR-associated proteins (CAS) were found in all strains except 006A-0059. The most striking difference in terms of genomic structure was the absence in commensal strains of several important genes involved in different steps of lipopolysaccharide biosynthesis (LPS) that were found in the pathogenic DSM 10053 (Supp. Fig. 6.2). In particular, commensal strains lacked heptosyltransferases involved in the biosynthesis of the inner core region. These enzymes are coded by *rfaC*, *rfaF* and *rfaQ* homologs in DSM 19053. The commensal strains also lacked *gmhABCD* homologs that are responsible for the transformation of D-sedoheptulose 7-phosphate into GDP-D-glycero-α-D-manno-heptose, an important constitutive of LPS core region. The O-antigen ligase RfaL (WaaL) coding gene was present only in DSM 19053, this ligase is crucial for the attachment of the O-antigen to the lipid A region. The presence of genes involved in the metabolism of sugars that constitute the O-antigen was also unpaired, showing great variability among commensal strains (Tab. **??**).

Table 6.1: Information for Canadian samples sequenced in this work.

Strain	Farm	Isolation date	Source	Material	ST
006A-0059	B25	04/04/06	dairy cattle	feces	U-tkt
006A-0063	B24	04/05/06	dairy cattle	feces	U-comb
006A-0073	B37	09/08/06	dairy cattle	feces	29
006A-0091	B52	31/01/07	dairy cattle	feces	U-aspA
006A-0113	B48	17/01/07	dairy cattle	feces	U-aspA
006A-0161	B77	08/08/07	beef cattle	feces	6
006A-0170	B79	08/08/07	beef cattle	feces	11
006A-0178	B80	08/08/07	dairy cattle	feces	U-comb
006A-0180	B80	08/08/07	dairy cattle	feces	U-glnA-pgm
006A-0191	B82	07/11/07	dairy cattle	feces	U-tkt
006A-0193	B85	07/11/07	beef cattle	feces	108
006A-0196	B87	28/11/07	dairy cattle	feces	123
007A-0283	-	18/07/05	environment	fresh water	48

6.4.2 Population structure and transmission. To study the genetic structure of this geographically confined population, we first built a core-genome phylogeny based on non-recombined regions (clonal frame) to look for signals of recent divergence and population structuring with BAPS. Fig. 6.1C shows that strains are accommodated in three different genomic lineages; also based on this analysis the pathogenic DSM 19053 belonged to a fourth genomic lineage distinct from commensal strains (Fig. 6.3).

To investigate the spatial structuring (Fig. 6.1A) in the population we built a linear regression using pairwise geographic distances (measured in km) and genetic distance (measured as number of mutations in the clonal frame) as variables. The lack of correlation between geographic and genetic distances (R^2 = 0.004, p-value = 0.25) (Fig. 6.4A) indicated that local differentiation driven by microevolutionary processes is not taking place in this *C. hyointestinalis* population. In addition, no correlation was found between genetic distance and isolation time measured in days (R^2 = 0.013, p-value = 0.96) (Fig. 6.4B). The lack of association between these variables suggests a high rate of spread of genetically diverse strains between farms. This is further supported by the fact that strain 006A-0180 (lineage 1) and strain 006A-0178 (lineage 3) were sampled on the same day on the same farm (B80), demonstrating the spatial co-occurrence of bacteria belonging to different lineages. The co-occurrence of the three genomic lineages identified here was supported by strain 006A-0161 (lineage 2), which was sampled in farm B77 the same day than the two previously mentioned strains.

Our results also demonstrate transmission of *C. hyointestinalis* between the environment and farms. This is shown by the close phylogenetic relationship between strains 007A-283 (isolated from an environmental freshwater course in July 2005) and 006A-0073 (isolated from daily cattle at farm B37 in August 2006). These two strains shared an immediate common ancestor, belonged to the same genomic lineage and were the pair of genomes with the lowest number of substitutions (195).

6.4.3 Genome-wide recombination and mutation rates. To identify the evolutionary forces responsible for the high genetic diversity observed in *C. hyointestinalis* we first applied the approach designed

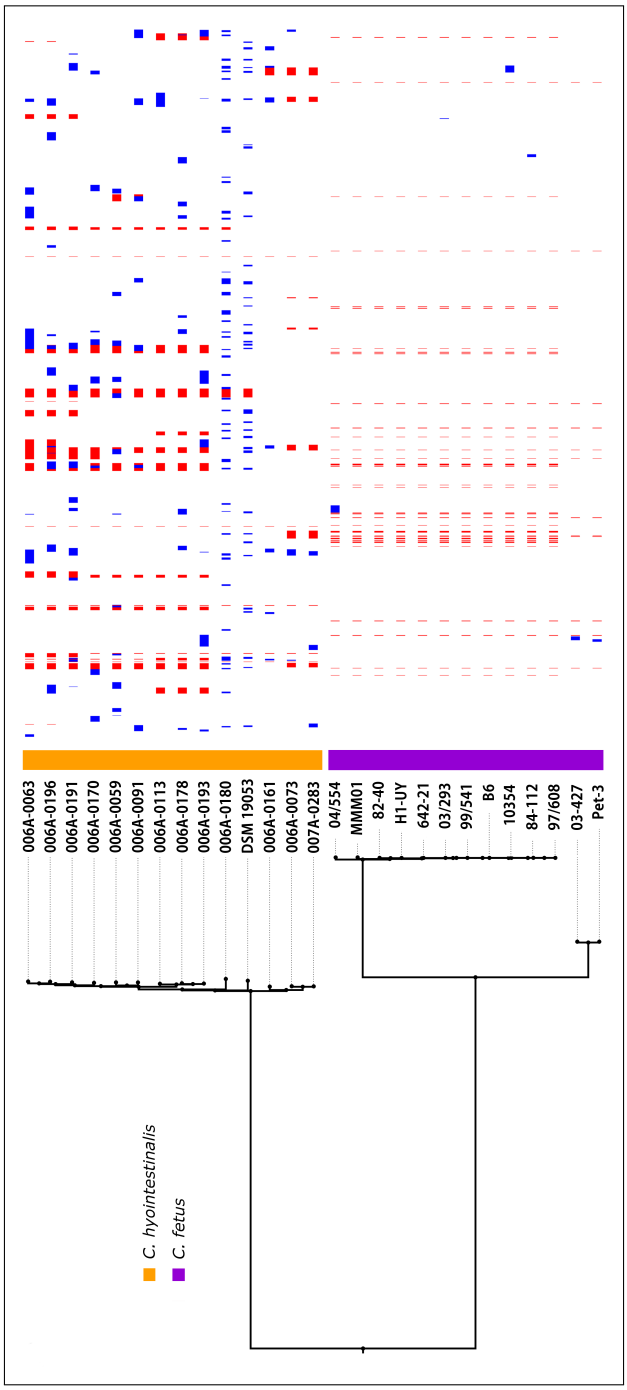


Figure 6.2: Recombination rates. Recombination analysis performed with Gubbins. Red blocks correspond to ancestrally shared recombination events while blue blocks are strain-specific recombinations.

by Croucher *et al.* [232] for detecting genome-wide recombination events. This algorithm is based on the identification of anomalous distributions of mutations along the genome (currently implemented in Gubbins software [229]). To observe and compare this phenomenon in a suitable phylogenetic and evolutionary framework, we inferred recombination along the *C. hyointestinalis* branch and its sister species *C. fetus* (Table S3), which share an immediate common ancestor. Fig. 6.2 clearly shows that ancestrally shared recombinations (red blocks) are more frequent and bigger across *C. hyointestinalis* branches than in *C. fetus* branches. Also, the incidence of strain-specific recombinations (blue blocks) is much greater in *C. hyointestinalis*. In general, all parameters used for measuring recombination were higher in *C. hyointestinalis* compared with *C. fetus*. In particular, the rate of recombination over mutation (r/m) was around 0.05 for all *C. fetus* while for *C. hyointestinalis* it was higher than 1.5 in some strains, however the wide dispersion in the distributions suggests a non-homogeneous effect of recombination along all *C. hyointestinalis* branches (Fig. 6.5A). The substitution rate inferred from the clonal frame (non-recombinant regions) in *C. fetus* genomes was 4.7×10^{-5} s/s/y (2.4×10^{-5}-5.1×10^{-6}, 95% HPD) while it was 1.4×10^{-3} s/s/y for *C. hyointestinalis* (1.2×10^{-3}- 4.3×10^{-3}, 95% HPD); this can be translated to an average of 50 expected fixed mutations between any pair of *C. fetus* genomes in one year, while this values increases to 1697 for *C. hyointestinalis* genomes (Fig. 6.5B).

A differential incidence of recombination was traceable to each genomic lineage, since each phylogenetic lineage displays a distinct recombination pattern. For example, lineage 2 was the least recombinogenic with a mean r/m value of 0.3 and lineage 3 was the most recombinogenic with a mean r/m value of 0.9 across its nodes (Tab. 6.2). Lineage 1 presented the highest number of non-ancestral (strain-specific) recombination, but also a high amount of mutations outside recombination that resulted in a relatively low ratio of recombination over mutations (r/m value of 0.5). In this analysis the genome of the reference strain DSM 19053 was included and also showed a high amount of non-ancestral recombination. The presence of intra-lineage variability was also observed, since some recombination events were not conserved in all members of lineage 3 and lineage 2.

Table 6.2: Recombination statistics expressed as intra-lineage means.

Lineage	r/m	Recombined bases	Clonal bases	Recombined SNPs	Clonal SNPs
1	0.5	29289	420214	1066	1228
2	0.3	16358	419109	172	491
3	0.9	46773	393919	265	361

Functional annotation was performed over those genes found inside lineage-specific recombinations and then this information was mapped onto KEGG metabolic pathways. Genes coding for thiamine metabolism (vitamin B1) were recombined in the three lineages, while genes for tricarboxylic acid cycle (TCA cycle), amino acids biosynthesis and pyrimidine metabolism were recombined in lineages 2 and 3 (Fig. 6.3A). To reveal associations between population structure and the acquisition or loss of particular set of genes across lineages, the pairwise Jaccard index calculated from presence/absence patterns of accessory genes and the pairwise number of substitutions in the core genome were plotted (Fig. 6.3B). This analysis revealed a positive significant correlation (R^2 = 0.4, p-value = 3×10^{-7}) between the genetic distance (measured as number of clonal substitutions) and the dissimilarity between accessory genes repertories (measured as Jaccard index). Indeed, the presence of lineage-specific genes was remarkable in lineage 1 (the longest branch in the tree) which presented 161 unique genes (Fig. 6.3C). Besides, members of linage 3 lacked *hsdR* and/or *hsdM* genes that codes for type I restriction-modification systems (R-M systems), while all members of lineage 1 and 2 contain at least one copy of each gene. Phylogenetic evidence supported that *hsdR* copies present in strains 006A-0180 (lineage 1) and 006A-0170 (lineage 3) were introduced by means of independent horizontal gene transfers (Supp. Fig. 6.6).

6.5 Discussion

Here we reveal the high rates of genetic diversity and the likely microevolutionary processes occurring in a natural population of *C. hyointestinalis*. Our observations are consistent with previous findings based on single genes or electrophoretic profiles that revealed great intra-specific diversity among *C. hyointestinalis* strains [212, 214, 233].

For other campylobacters, like *C. jejuni* and *C. fetus*, the presence of

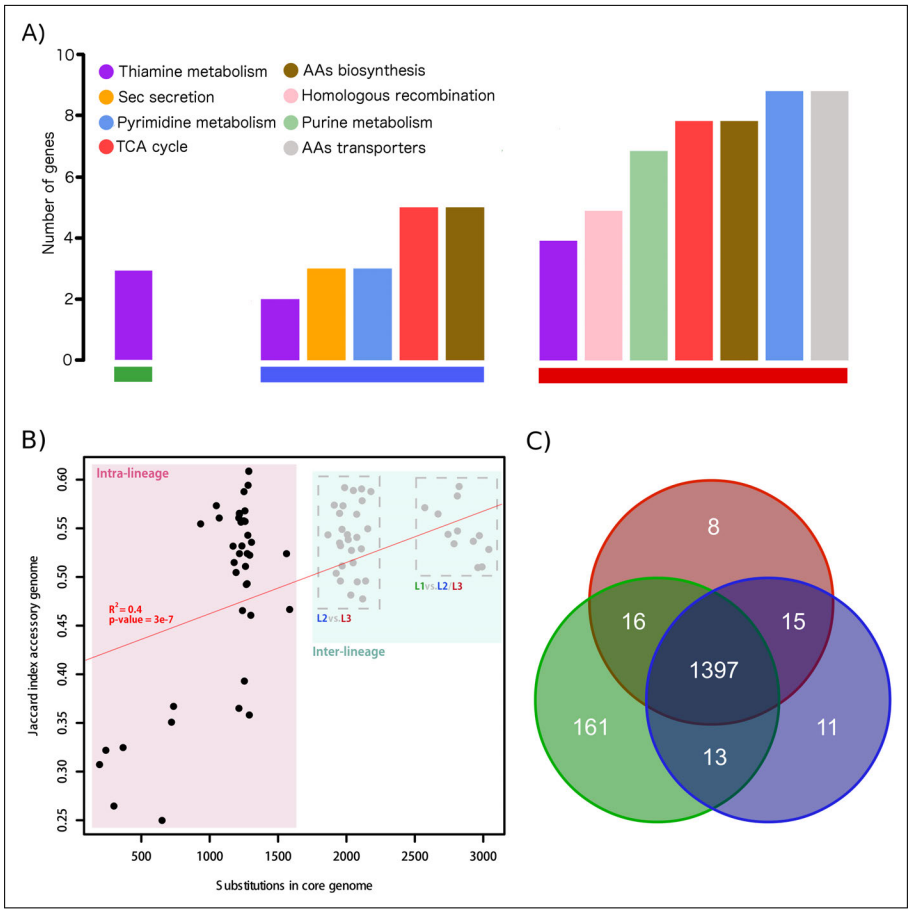


Figure 6.3: Recombination and pan-genome. A) Functional annotation of genes inside lineage-specific recombinations. B) Correlation between genetic distance measured as number of substitutions in the clonal frame versus the Jaccard dissimilarity index from accessory genes. C) Venn diagram showing the number of exclusive genes present in each intersection between lineages.

host-associated STs that transcends geographic variation has been demonstrated [140, 234], however this was not known for *C. hyointestinalis*. The surprising heterogeneity found among strains from very close locations suggest the absence of a cattle-associated genotype and also highlight the genomic plasticity of this species that probably accounts for its high adaptive potential. The heterogeneity was also evident when looking at general genomic features: in contrast with a

general trend for the genus *Campylobacter* that indicate an overall GC variation below 1% for genomes belonging to the same species, *C. hyointestinalis* genomes varied about 3%. This may reflect a speciation process occurring in *C. hyointestinalis* populations, supported by the low ANI values (<95%) observed between commensal strains and the pathogenic DSM 19053. Some other notable differences were present in the LPS biosynthesis locus. In particular, the absence of key genes for its constitutive components in commensal strains with respect to the pathogenic DSM 19053 may indicate the importance of LPS in *C. hyointestinalis* virulence, like has been demonstrated for other gram-negatives such as *Salmonella* [235] and *Francisella* [236]. However, this evidence should be tested using experimental procedures and a larger collection of isolates from healthy and diseased animals.

The tremendous diversity within the population is driven by both recombination and clonal diversification. The ratio of recombination over mutation (r/m) is used to measure the impact of these evolutionary forces, varying in orders of magnitude depending on the considered species [237]. In *C. hyointestinalis* r/m was not as high as in other recombinogenic bacteria like *Neisseria gonorrhoeae* [238], however this was caused by an extremely high substitution rate instead of low recombination rate. Indeed, a substitution rate has been recently estimated around 5×10^{-5} s/s/y for some clonal complexes of *C. jejuni* [239], which is similar to our estimates from public genomes of *C. fetus*, far lower than the rapidly evolving genomes of *C. hyointestinalis*. It is worth mentioning that as recombination events are likely to introduce several substitutions, a r/m ~1 will usually produce a greater per-site effect of recombination than mutation.

Despite the high evolutionary rates a clear phylogenetic structure was evidenced where each lineage presented differential incidence of recombination, linked to the presence/absence of Restriction-Modification systems (R-M systems) which tend to undermine the effects of recombination by preventing DNA integration [240]. This finding suggests a correlation between the clonal diversification of the population and gene gain/loss events. A similar phenomenon has been recently described in *C. jejuni* where distinctive patterns of intra-lineage recombination were accompanied by the presence of lineage-specific R-M systems [241]. Recombinant regions in *C. hyointestinalis* harboured

genes of basal biological processes like nucleotides metabolism. Interestingly, high variation at purine biosynthesis genes are associated to enhanced fitness under stress response in *C. jejuni* [242]. By extrapolating this to *C. hyointestinalis*, recombination on housekeeping loci could provide a source of diversity that favours host adaptation and transmission.

Taken together our results represent the first genomic analysis of commensal *C. hyointestinalis*, uncovering the main forces shaping its evolution and providing evidence for the transmission mechanisms of this emerging pathogen. Remarkably, we revealed for the first time the potential of transmission of *C. hyointestinalis* between farms and the environment. Other species like *C. jejuni*, *C. coli* and *C. lari* are also able to persist in the environment [243] and transmit to different hosts, hence *C. hyointestinalis* could have a similar ecological behaviour which not only explains its wide host range but also its role as commensal and opportunistic pathogen. Undoubtedly, the extension of the current genomic dataset by incorporating strains derived from other hosts can shed more light on the evolutionary dynamics of this species. Our analysis demonstrates that the genomic diversity of the *Campylobacter* genus is undersampled and that paying attention to non-classical campylobacters can uncover important epidemiological features of previously under studied species.

6.6 Acknowledgements

We thank the Pathogen Core Informatics group at the Wellcome Trust Sanger Institute for valuable technical assistance.

6.7 References

[79] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **1990**, *215*, 403–410.

[140] M. A. van Bergen, K. E. Dingle, M. C. Maiden, D. G. Newell, L. van der Graaf-Van Bloois, J. P. van Putten, J. A. Wagenaar, *J. Clin. Microbiol.* **2005**, *43*, 5888–5898.

[145] S. M. Man, *Nat Rev Gastroenterol Hepatol* **2011**, *8*, 669–685.

[149] D. R. Zerbino, E. Birney, *Genome Res.* **2008**, *18*, 821–829.

[208] C. Friedman, J. Neimann, H. Wegener, R. Tauxe in *Campylobacter*, *Vol. II/6*, ASM International, **2000**, pp. 121–138.

[209] C. J. Gebhart, G. E. Ward, K. Chang, H. J. Kurtz, *Am. J. Vet. Res.* **1983**, *44*, 361–367.

[210] G. Gorkiewicz, G. Feierl, R. Zechner, E. L. Zechner, *J. Clin. Microbiol.* **2002**, *40*, 2601–2605.

[211] S. Bullman, D. Corcoran, J. O'Leary, D. O'Hare, B. Lucey, R. D. Sleator, *FEMS Immunol. Med. Microbiol.* **2011**, *63*, 248–253.

[212] S. L. On, M. Costas, B. Holmes, *Systematic and Applied Microbiology* **1993**, *16*, 37 –46.

[213] W. G. Miller, M. H. Chapman, E. Yee, S. L. On, D. K. McNulty, A. J. Lastovica, A. M. Carroll, E. B. McNamara, G. Duffy, R. E. Mandrell, *Front Cell Infect Microbiol* **2012**, *2*, 45.

[214] C. S. Harrington, S. L. On, *Int. J. Syst. Bacteriol.* **1999**, *49 Pt 3*, 1171–1175.

[215] S. M. Salama, H. Tabor, M. Richter, D. E. Taylor, *J. Clin. Microbiol.* **1992**, *30*, 1982–1984.

[216] M. He, F. Miyajima, P. Roberts, L. Ellison, D. J. Pickard, M. J. Martin, T. R. Connor, S. R. Harris, D. Fairley, K. B. Bamford, S. D'Arc, J. Brazier, D. Brown, J. E. Coia, G. Douce, D. Gerding, H. J. Kim, T. H. Koh, H. Kato, M. Senoh, T. Louie, S. Michell, E. Butt, S. J. Peacock, N. M. Brown, T. Riley, G. Songer, M. Wilcox, M. Pirmohamed, E. Kuijper, P. Hawkey, B. W. Wren, G. Dougan, J. Parkhill, T. D. Lawley, *Nat. Genet.* **2013**, *45*, 109–113.

[217] M. Boetzer, C. V. Henkel, H. J. Jansen, D. Butler, W. Pirovano, *Bioinformatics* **2011**, *27*, 578–579.

[218] M. Boetzer, W. Pirovano, *Genome Biol.* **2012**, *13*, R56.

[219] T. Seemann, *Bioinformatics* **2014**, *30*, 2068–2069.

[220] K. Tamura, G. Stecher, D. Peterson, A. Filipski, S. Kumar, *Mol. Biol. Evol.* **2013**, *30*, 2725–2729.

[221] S. Sunagawa, D. R. Mende, G. Zeller, F. Izquierdo-Carrasco, S. A. Berger, J. R. Kultima, L. P. Coelho, M. Arumugam, J. Tap, H. B. Nielsen, S. Rasmussen, S. Brunak, O. Pedersen, F. Guarner, W. M. de Vos, J. Wang, J. Li, J. Dore, S. D. Ehrlich, A. Stamatakis, P. Bork, *Nat. Methods* **2013**, *10*, 1196–1199.

[222] K. T. Konstantinidis, J. M. Tiedje, *J. Bacteriol.* **2005**, *187*, 6258–6264.

[223] N. J. Croucher, A. J. Page, T. R. Connor, A. J. Delaney, J. A. Keane, S. D. Bentley, J. Parkhill, S. R. Harris, *Nucleic Acids Res.* **2015**, *43*, e15.

[224] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, M. A. Marra, *Genome Res.* **2009**, *19*, 1639–1645.

[225] A. J. Page, C. A. Cummins, M. Hunt, V. K. Wong, S. Reuter, M. T. Holden, M. Fookes, D. Falush, J. A. Keane, J. Parkhill, *Bioinformatics* **2015**, *31*, 3691–3693.

[226] A. Loytynoja, N. Goldman, *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 10557–10562.

[227] A. Stamatakis, *Bioinformatics* **2014**, *30*, 1312–1313.

[228] J. Corander, P. Marttinen, J. Siren, J. Tang, *BMC Bioinformatics* **2008**, *9*, 539.

[229] N. J. Croucher, A. J. Page, T. R. Connor, A. J. Delaney, J. A. Keane, S. D. Bentley, J. Parkhill, S. R. Harris, *Nucleic acids research* **2014**, gku1196.

[230] Y. Moriya, M. Itoh, S. Okuda, A. C. Yoshizawa, M. Kanehisa, *Nucleic Acids Res.* **2007**, *35*, W182–185.

[231] R. Bouckaert, J. Heled, D. Kuhnert, T. Vaughan, C. H. Wu, D. Xie, M. A. Suchard, A. Rambaut, A. J. Drummond, *PLoS Comput. Biol.* **2014**, *10*, e1003537.

[232] N. J. Croucher, S. R. Harris, C. Fraser, M. A. Quail, J. Burton, M. van der Linden, L. McGee, A. von Gottberg, J. H. Song, K. S. Ko, et al., *Science* **2011**, *331*, 430–434.

[233] S. M. Salama, M. M. Garcia, D. E. Taylor, *International Journal of Systematic and Evolutionary Microbiology* **1992**, *42*, 446–450.

[234] S. K. Sheppard, F. Colles, J. Richardson, A. J. Cody, R. Elson, A. Lawson, G. Brick, R. Meldrum, C. L. Little, R. J. Owen, et al., *Applied and Environmental Microbiology* **2010**, *76*, 5269–5277.

[235] Q. Kong, J. Yang, Q. Liu, P. Alamuri, K. L. Roland, R. Curtiss, *Infection and immunity* **2011**, *79*, 4227–4239.

[236] T.-H. Kim, J. T. Pinkham, S. J. Heninger, S. Chalabaev, D. L. Kasper, *Journal of Infectious Diseases* **2011**, jir620.

[237] X. Didelot, M. C. Maiden, *Trends in microbiology* **2010**, *18*, 315–322.

[238] M. Pérez-Losada, E. B. Browne, A. Madsen, T. Wirth, R. P. Viscidi, K. A. Crandall, *Infection Genetics and Evolution* **2006**, *6*, 97–112.

[239] B. L. Dearlove, A. J. Cody, B. Pascoe, G. Méric, D. J. Wilson, S. K. Sheppard, *The ISME journal* **2015**.

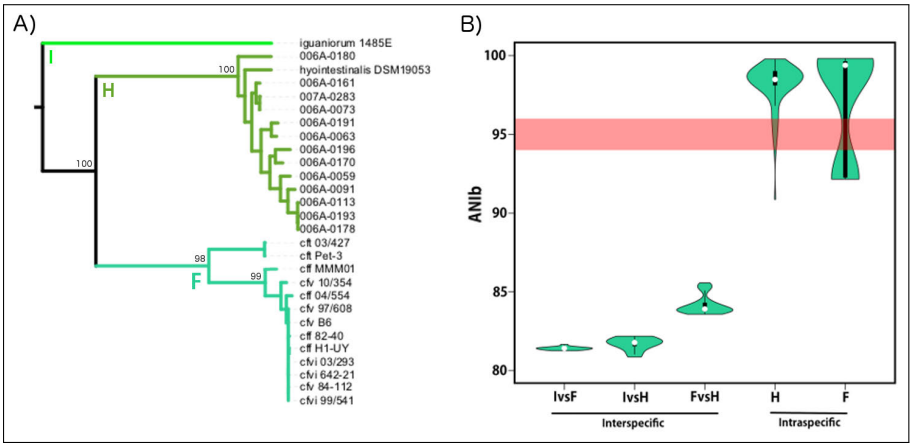
[240] K. Vasu, V. Nagaraja, *Microbiology and Molecular Biology Reviews* **2013**, *77*, 53–72.

[241] L. Morley, A. McNally, K. Paszkiewicz, J. Corander, G. Méric, S. K. Sheppard, J. Blom, G. Manning, *Applied and environmental microbiology* **2015**, *81*, 3641–3647.

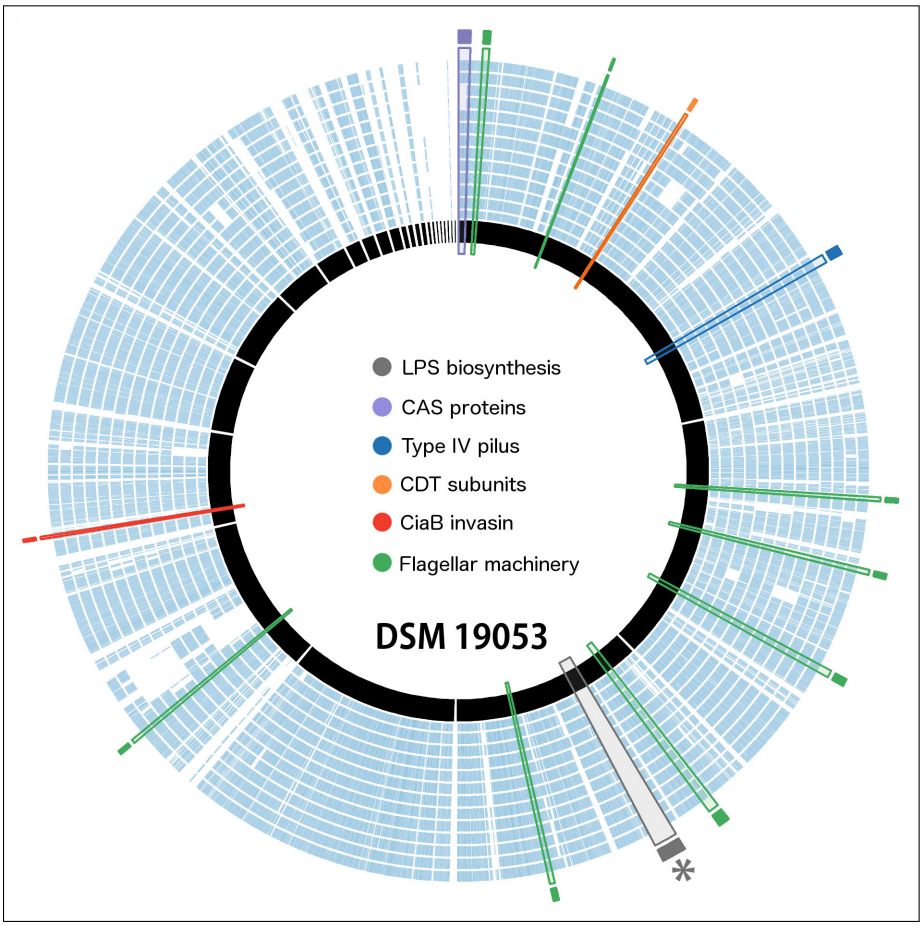
[242] A. Cameron, S. Huynh, N. E. Scott, E. Frirdich, D. Apel, L. J. Foster, C. T. Parker, E. C. Gaynor, *mBio* **2015**, *6*, e00612–15.

[243] K. Jones, *Journal of Applied Microbiology* **2001**, *90*.

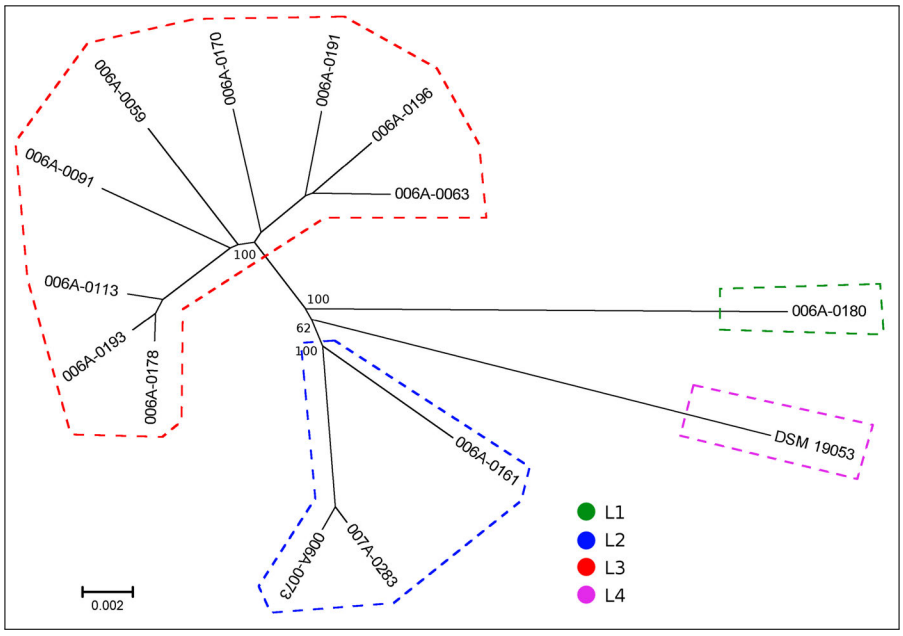
6.8 Supplementary material



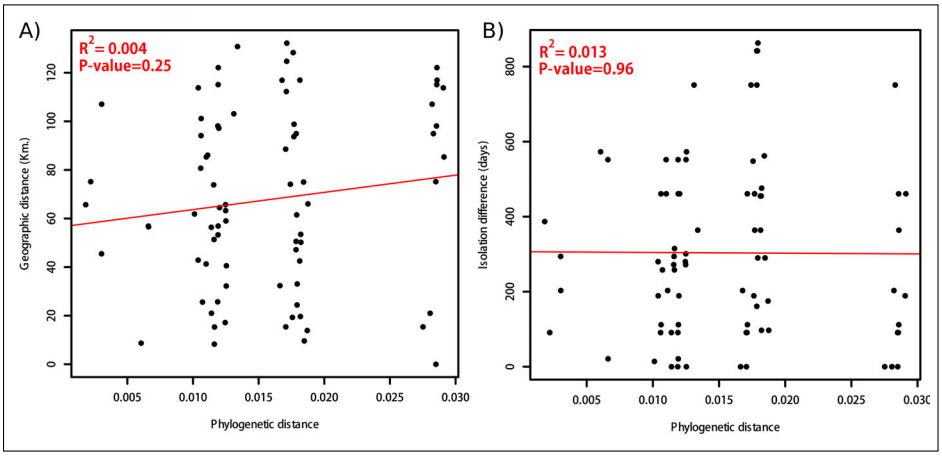
Supp. Fig. 6.1: Species phylogeny and Average Nucleotide Identity calculations. A) Maximum Likelihood phylogeny built with the concatenated sequences of 40 prokaryotic universal proteins. Species last common ancestors are highlighted with capital letters: I) *C. iguaniorum*, H) *C. hyointestinalis* and F) *Campylobacter fetus*. Bootstrap values for relevant internals nodes are shown. The topology clearly supports the classification of the strains sequenced in this study as *C. hyointestinalis*. B) Violin plots showing the inter-specific and intra-specific distributions of Average Nucleotide Identity values calculated with BLAST (ANIb). The inter-specific distributions support the phylogeny since ANIb from genomes belonging to different species are far below 95%. The intraspecific distribution for *C. hyointestinalis* (H) has a mean near 100% indicating that most genomes can be assigned to *C. hyointestinalis* species. The tail that falls below 95% is produced by low ANIb values resulting from comparisons between DSM 19053 (pathogenic) and the rest sequenced in this study (commensal).



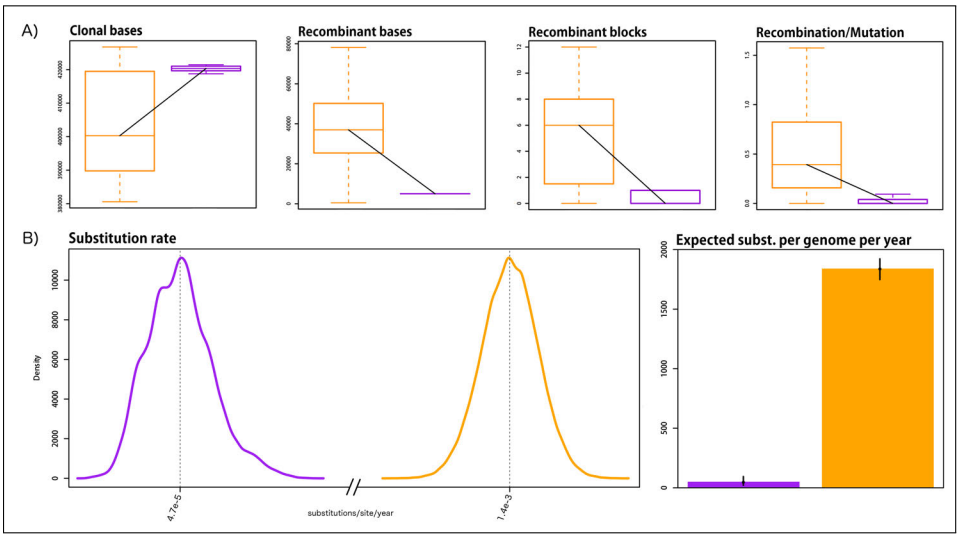
Supp. Fig. 6.2: Circos representation. Conserved genes between the pathogenic DSM 19053 (black inner ring) and each commensal *C. hyointestinalis* genome sequenced in this work (blue outer rings). The asterisk highlights the LPS region absent in commensal strains but present in the pathogenic DSM 19053.



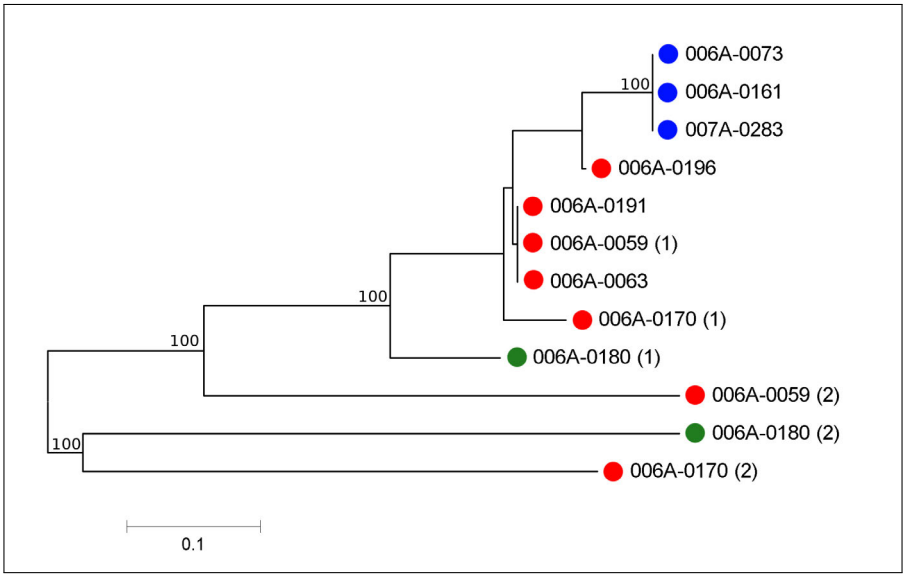
Supp. Fig. 6.3: Whole-genome phylogeny. Maximum Likelihood phylogeny using concatenated core genome genes without recombinant regions. The reference strain DSM 19053 is highlighted as a fourth genomic linage based on BAPS analysis. Bootstrap values are indicated for relevant internal nodes.



Supp. Fig. 6.4: Geographic and temporal correlations. A) Linear regression between geographic distance measured in km and phylogenetic distance measured as number of clonal mutations. B) Linear regression between isolation time difference (measured in days) and phylogenetic distance measured as number of clonal mutations.



Supp. Fig. 6.5: Recombination and mutation rates. A) Box plots showing the number of clonal bases, recombinant bases, recombinant blocks and r/m along *C. hyointestinalis* (orange) and *C. fetus* (purple) nodes. B) Posterior probability densities and bar plots for substitution rates in *C. hyointestinalis* and *C. fetus* clonal frames.



Supp. Fig. 6.6: Phylogeny of *hsdR* genes. Maximum likelihood tree, strains are colored by lineage: lineage 1 in green, lineage 2 in blue and lineage 3 in red. When multiple copies occur per genome this is indicated inside brackets.

Supp. Tab. 6.1: Distribution of LPS genes in *C. hyointestinalis* genomes.

KEGG Orthology	Description	Region	Gene name	006A-0180	007A-0283	006A-0073	006A-0161	006A-0059	006A-0063	006A-0091	006A-0113	006A-0170	006A-0178	006A-0191	006A-0193	006A-0196	DSM 19053
K00996	Undecaprenyl-phosphate galactose phosphotransferase	O-antigen	rfbP;wcaJ;weaP;wbaP;amsP	2	2	2	2	2	2	2	2	2	2	2	2	2	2
K02847	Lipid A core - O-antigen ligase and related enzymes	O-antigen	waaL;rfaL	0	0	0	0	0	0	0	0	0	0	0	0	0	1
K02851	UDP-GlcNAc:undecaprenyl-phosphate transferase	O-antigen	wecA;tagO;rfe	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K05789	chain length determinant protein	O-antigen	wzzB;cld	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K07272	rhamnosyltransferase	O-antigen	rgpF	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K08280	lipopolysaccharide O-acetyltransferase	O-antigen	wbbJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K12990	rhamnosyltransferase	O-antigen	rfbF	2	2	2	1	1	1	1	1	1	1	1	1	1	1
K12991	rhamnosyltransferase	O-antigen	rfbG	1	1	1	0	0	1	1	1	1	1	1	1	1	1
K12992	rhamnosyltransferase	O-antigen	rfbN	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K12993	rhamnosyltransferase	O-antigen	wbpX	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K12994	alpha-1,3-rhamnosyl/mannosyltransferase	O-antigen	wbpT;wbdB	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K12995	rhamnosyl/mannosyltransferase	O-antigen	wbpZ;wbdC	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K12996	rhamnosyltransferase	O-antigen	rgpA	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K12997	rhamnosyltransferase	O-antigen	rgpB;opsA	0	0	0	0	0	0	1	1	1	1	0	1	0	0
K12998	glucosyltransferase	O-antigen	rgpE	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K12999	glucosyltransferase	O-antigen	rgpI	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K13000	mannosyltransferase	O-antigen	wbyJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K13001	mannosyltransferase	O-antigen	wbyK	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K13002	glycosyltransferase	O-antigen	wbyL	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K13003	glycosyltransferase	O-antigen	wbtG	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K13004	galacturonosyltransferase	O-antigen	wbtD	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K13005	abequosyltransferase	O-antigen	rfbV	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K13007	Fuc2NAc and GlcNAc transferase	O-antigen	wbpL	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K13008	O-antigen polymerase	O-antigen	rfc;wbbH;wzyB	1	1	1	1	1	1	1	1	1	1	1	1	1	0
K13009	O-antigen polymerase	O-antigen	wzy	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K13011	O-antigen biosynthesis protein WbqL	O-antigen	wbqL	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K13012	O-antigen biosynthesis protein WbqP	O-antigen	wbqP	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K18799	O-antigen flippase	O-antigen	wzx;rfbX	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K18827	O-antigen chain-termination methyltransferase	O-antigen	wbbD	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K02527	3-deoxy-D-manno-octulosonic-acid transferase	Core region	waaA	1	1	1	1	1	1	1	1	1	1	1	1	1	1
K12981	KDO transferase III	Core region	waaZ;rfaZ	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K02841	lipopolysaccharide heptosyltransferase I	Core region	waaC;rfaC	0	0	0	0	0	0	0	0	0	0	0	0	0	1
K02843	lipopolysaccharide heptosyltransferase II	Core region	waaF;rfaF	0	0	0	0	0	0	0	0	0	0	0	0	0	1
K02849	putative heptosyltransferase III waaq	Core region	waaQ;rfaQ_1	0	0	0	0	0	0	0	0	0	0	0	0	0	2
K03277	heptosyltransferase IV	Core region	waaU;rfaK	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K12982	heptosyltransferase I	Core region	opsX	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K02844	UDP-glucose LPS alpha-1,3-glucosyltransferase	Core region	rfaG;wabG	1	1	1	1	1	1	1	1	1	1	1	1	1	1
K03275	UDP-glucose LPS alpha-1,3-glucosyltransferase	Core region	waaO;rfaI	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K03276	UDP-glucose LPS alpha-1,3-glucosyltransferase	Core region	waaR;waaT;rfaJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K03279	UDP-glucose LPS alpha-1,3-glucosyltransferase	Core region	waaJ;rfaJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K12983	UDP-glucose LPS beta-1,3-glucosyltransferase	Core region	waaV	1	1	1	0	0	0	0	0	0	0	0	0	0	0
K12984	UDP-glucose LPS beta-1,3-glucosyltransferase	Core region	waaE;kdtX	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K02840	UDP-glucose LPS alpha-1,6-galactosyltransferase	Core region	waaB;rfaB	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K03278	UDP-galactose LPS alpha-1,3-D-galactosyltransferase	Core region	waaI;rfaI	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K12985	LPS 1,2-glucosyltransferase	Core region	waaW	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K03280	UDP-N-acetylglucosaminetransferase	Core region	waaK;rfaK	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K12986	1,5-rhamnosyltransferase	Core region	waaS;rfaS	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K12987	alpha-1,6-rhamnosyltransferase	Core region	migA	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K12988	alpha-1,3-rhamnosyltransferase	Core region	wapR	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K12989	mannosyltransferase	Core region	lpcC	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K02848	heptose I phosphotransferase	Core region	waaP;rfaP	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K02850	heptose II phosphotransferase	Core region	waaY;rfaY	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K12975	KDO II ethanolaminephosphptransferase	Core region	eptB	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K19353	heptose I phosphate ethanolaminephosphotransferase	Core region	eptC	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K00713	UDP-glucose LPS alpha-1,2-glucosyltransferase	Core region	waaD	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K19354	heptose III glucuronosyltransferase	Core region	waaH	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K06041	arabinose-5-phosphate isomerase	Unusual sugars	kdsD;kpsF	1	1	1	1	1	1	1	1	1	1	1	1	1	2
K01627	2-dehydro-3-deoxyphosphooctonate aldolase	Unusual sugars	kdsA	1	1	1	1	1	1	1	1	1	1	1	1	1	1
K03270	3-deoxy-D-manno-octulosonate 8-phosphate phosphatase	Unusual sugars	kdsC	1	1	1	1	1	1	1	1	1	1	1	1	1	1
K00979	3-deoxy-manno-octulosonate cytidylyltransferase	Unusual sugars	kdsB	1	1	1	1	1	1	1	2	2	2	2	2	1	2
K03271	D-sedoheptulose 7-phosphate isomerase	Unusual sugars	gmhA	0	0	0	0	0	0	0	0	0	0	0	0	0	1
K03272	D-beta-D-heptose 7-phosphate kinase	Unusual sugars	gmhC;hldE	0	0	0	0	0	0	0	0	0	0	0	0	0	1
K03273	D-glycero-D-manno-heptose 1,7-bisphosphate phosphatase	Unusual sugars	gmhB	0	0	0	0	0	0	0	0	0	0	0	0	0	1
K03274	ADP-L-glycero-D-manno-heptose 6-epimerase	Unusual sugars	gmhD;rfaD	0	0	0	0	0	0	0	0	0	0	0	0	0	1
K10011	UDP-4-amino-4-deoxy-L-arabinose formyltransferase	Unusual sugars	arnA;pmrI	2	2	2	2	2	2	3	3	3	3	2	3	2	3
K07806	UDP-4-amino-4-deoxyl-L-arabinose-oxoglutarate aminotransferase	Unusual sugars	arnB;pmrH	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K10012	undecaprenyl-phosphate 4-deoxy-4-formamido-L-arabinose transferase	Unusual sugars	arnC;pmrF	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K13014	undecaprenyl phosphate-alpha-L-ara4FN deformylase	Unusual sugars	arnD	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K12962	undecaprenyl phosphate-alpha-L-ara4N flippase subunit ArnE	Unusual sugars	arnE	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K12963	undecaprenyl phosphate-alpha-L-ara4FN flippase subunit ArnF	Unusual sugars	arnF	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K13015	UDP-N-acetyl-D-glucosamine dehydrogenase	Unusual sugars	wbpA	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K02474	UDP-N-acetyl-D-galactosamine dehydrogenase	Unusual sugars	wbpO	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K13016	UDP-N-acetly-2-amino-2-deoxyglucuronate dehydrogenase	Unusual sugars	wbpB	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K13020	UDP-N-acetly-2-amino-2-deoxyglucuronate dehydrogenase	Unusual sugars	wlbA;lpsIA	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K13017	UDP-2-acetamido-2-deoxy-ribo-hexuluronate aminotransferase	Unusual sugars	wbpE;wlbC	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K13018	UDP-2-acetamido-3-amino-2,3-dideoxy-glucuronate N-acetyltransferase	Unusual sugars	wbpD;wlbB	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K01791	UDP-N-acetylglucosamine 2-epimerase	Unusual sugars	wecB	0	0	0	1	0	1	1	1	1	1	0	1	1	0
K13019	UDP-GlcNAc3NAcA epimerase	Unusual sugars	wbpI	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K13006	UDP-perosamine 4-acetyltransferase	Unusual sugars	wbqR	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K13010	perosamine synthetase	Unusual sugars	per;rfbE	1	1	1	1	1	1	1	1	1	1	1	1	1	1
K17939	GDP-perosamine N-acetyltransferase	Unusual sugars	perB;wbdR	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K13013	O-antigen biosynthesis protein WbqV	Unusual sugars	wbqV	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K00748	Lipid-A-disaccharide synthase	Lipid A	lpxB	1	1	1	1	1	1	1	1	1	1	1	1	1	1
K00677	UDP-N-acetylglucosamine acyltransferase	Lipid A	lpxA	1	1	1	1	1	1	1	1	1	1	1	1	1	1
K02536	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase	Lipid A	lpxD	1	1	1	1	1	1	1	1	1	1	1	1	1	1
K02517	KDO2-lipid IV(A) lauroyltransferase	Lipid A	lpxL;htrB	1	1	1	1	1	1	1	1	1	1	1	1	1	1
K02560	lauroyl-KDO2-lipid IV(A) myristoyltatransferase	Lipid A	lpxM;msbB	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K12973	palmitoyl transferase	Lipid A	pagP	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K12974	KDO2-lipid IV(A) palmitoleoyltransferase	Lipid A	lpxP	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K03760	lipid A ethanolaminephosphotransferase	Lipid A	eptA	1	1	1	1	1	1	1	1	1	1	1	1	1	1
K02535	UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase	Lipid A	lpxC	1	1	1	1	1	1	1	1	1	1	1	1	1	1
K16363	(3R)-hydroxymyristoyl-ACP dehydratase	Lipid A	fabZ	1	1	1	1	1	1	1	1	1	1	1	1	1	1
K12976	lipid A 3-O-deacylase	Lipid A	pagL	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K09953	lipid A 3-O-deacylase	Lipid A	lpxR	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K03269	UDP-2,3-diacylglucosamine hydrolase	Lipid A	lpxH	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K07264	4-amino-4-deoxy-L-arabinose transferase	Lipid A	arnT;pmrK	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K00912	tetraacyldisaccharide 4′-kinase	Lipid A	lpxK	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K12977	lipid A 1-phosphatase	Lipid A	lpxE	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K12978	lipid A 4′-phosphatase	Lipid A	lpxF	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K12979	Beta-hydroxylase	Lipid A	lpxO	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K12980	lipid A oxidase	Lipid A	lpxQ	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Conclusión del Capítulo 6

Campylobacter hyointestinalis es una especie sumamente poco estudiada desde el punto de vista genómico, a pesar que sus características biológicas la hacen un modelo realmente atractivo; en particular en un par de estudios pioneros basados en diversidad de los genes ribosomales se concluyó que esta especie posee una alta tasa de mutación impropia de sus parientes más cercanos. Con respecto a su importancia sanitaria, solo recientemente se ha identificado su potencial zoonótico y se han establecido conexiones epidemiológicas entre humanos y animales. Desde el punto de vista genómico no hay estudios que hayan analizado su plasticidad y evolución, habiendo disponible tan solo un genoma completo de una cepa aislada de un cerdo con enteritis proliferativa.

El principal aporte de este trabajo fue generar los primeros genomas de cepas de *C. hyointestinalis* aisladas de bovinos sin síntomas de enfermedad y del ambiente. Estos genomas fueron utilizados para estudiar la pasticidad genómica de la especie en el contexto de este hospedero.

Se determinó que *C. hyointestinalis* posee una tasa de recombinación muy superior a sus especies hermanas, además evoluciona por mutación a una tasa muy elevada. La información genómica también permitió postular la presencia de una estructuración poblacional en una regioń geográfica muy pequeña. Estos linajes poblacionales co-ocurren en las mismas granjas demostrando una alta tasa de transmisión de cepas genéticamente diferentes. Además la presencia de cepas aisladas de cursos de agua naturales sustenta su potencial subsistencia en el ambiente y transmisión hacia los animales. Basados en estos resultados, se propone que la gran pasticidad genómica de *C. hyointestinalis* es un factor importante para su gran adaptabilidad a los hospederos mamíferos, lo cual convierte a esta especie en un potencial patógeno emergente para animales de producción y humanos.

7

Campylobacter geochelonis: a new species from Hermann's testudines

Citation:

Piccirillo A*, Niero G, Calleros L, Pérez R, Naya H, Iraola G*. (2016) ***Campylobacter geochelonis*** **sp. nov., isolated from the western Hermann's tortoise (*Testudo hermanni hermanni*)** . *International Journal of Systematic and Evolutionary Microbiology*. Accepted.

* Corresponding authors

7.1 Abstract

During a screening study aimed at assessing the presence of *Campylobacter* spp. in reptiles, three putative strains (RC7, RC11, and RC20) were isolated from different individuals of the western Hermann's tortoise (*Testudo hermanni hermanni*). Initially, isolates were characterized as *C. fetus* subsp. *fetus* by multiplex PCR and partial 16S rRNA sequence analysis. Further whole genome characterization revealed considerable differences compared to other *Campylobacter* species. A polyphasic study was then undertaken to determine the exact taxonomic position of the isolates. The three strains were characterized by conventional phenotypic tests and whole-genome sequencing. We generated robust phylogenies that showed a distinct clade containing only these strains using the 16S rRNA and *atpA* genes and a set of 40 universal proteins. Our phylogenetic analysis demonstrates their designation as a new species and this was further confirmed using whole genome average nucleotide identity within the *Campylobacter* genus (~80%). Comparted to most *Campylobacter* species these strains hydrolysed hippurate, grew well at 25 and showed resistance to nalidixic acid. They grew well in blood agar at 25 °C but not at 42°C. Phenotypic and genetic analyses demonstrate that the three *Campylobacter* strains isolated from the western Hermann's tortoise represent a novel species within the *Campylobacter* genus, for which the name *Campylobacter geochelonis* sp. nov. is proposed, with RC20^T (= DSM 102159^T = LMG 29375^T) as the type strain.

7.2 Phenotypic and genomic characterization

The *Campylobacter* genus belongs to the family Campylobacteraceae, order Campylobacterales, class Epsilonproteobacteria [244]. To date, 34 species and 14 subspecies have been recognized within the genus (LPSN, 2016), with *C. fetus* as the type species [138]. In the last years, molecular biology techniques, mainly high throughput sequencing, have dramatically accelerated the identification of novel *Campylobacter* species [245]. *Campylobacter* spp. are commensal or pathogenic bacteria of a broad range of mammalian, avian and reptilian species [246]. Thermophilic *Campylobacter* spp. are associated mainly with poultry

and most of them have zoonotic potential [145]. To date, only two *Campylobacter* taxa of reptile origin have been identified: *C. fetus* subsp. *testudinum* and *C. iguaniorum* [202, 247]. *C. fetus* subsp. *testudinum* has been isolated from reptiles and humans [202]; *C. iguaniorum* have been isolated from lizards and chelonians (Gilbert et al., 2015). In this study, we describe the results of a polyphasic taxonomic study aimed at characterizing three *Campylobacter* strains isolated from western Hermann's tortoises (*Testudo hermanni hermanni*).

In 2011, a study aimed at assessing the presence of *Campylobacter* spp. in captive reptiles was carried out in Northern Italy [248]. Samples were collected from 11 turtle species belonging to four different families (Chelydridae, Emydidae, Testudinidae and Trionychidae). Animals were kept in a zoo and in private households in Northern Italy. Cloacal swabs were collected and transported to the laboratory in Amies with charcoal medium (Copan, Brescia, Italy), and immediately inoculated in Preston broth (Oxoid, Milano, Italy) and incubated at 37°C for 24 h in a microaerobic atmosphere (CampygenTM, Oxoid). Broth cultures were filtrated [249], plated onto Nutrient agar with 5% sheep blood (Oxoid) and incubated at 37°C for 96 h under microaerobic conditions. Among all the *Campylobacter* isolates, three strains from different and geographically unrelated turtles belonging to *Testudo hermanni hermanni* species were identified as *C. fetus* subsp. *fetus* by multiplex PCR [250], 16S rRNA sequencing [251], and the PCR protocol described by [252] specifically developed for *C. fetus* sub-typing. Strains were designated with the following reference names: RC7, RC11 and RC20; and stored at -80°C. Afterward, the whole genome of these strains was determined as part of a large-scale study on the genetic diversity of *C. fetus* and genome comparisons revealed that the classification as *C. fetus* was inconsistent, which was also supported by additional PCR sub-typing markers [86, 203] and partial 16S sequencing as described in Linton *et al.* [147]. In the present paper, we report the results of a polyphasic taxonomic study carried out to determine the exact classification of the strains. To this end, phenotypic and genotypic characteristics of the three *Campylobacter* strains were determined by classical biochemical testing and whole genome comparison.

To perform all the analyses, strains were grown on Tryptone Soya Agar (TSA) with 5% sheep blood (Oxoid) under microaerobic condi-

tions at 37°C for 48 h. Bacterial DNA was extracted by using the Invisorb®spin tissue mini kit (Stratec, Birkenfeld, Germany), following the manufacturer's instructions. Whole genome sequences of RC7, RC11 and RC20 strains were obtained with an Illumina HiSeq 2000 sequencer that produced 100 bp pair-end reads at an average coverage of 300x. Draft genomes were assembled and annotated using Velvet [149] and Prokka [219], respectively. The raw reads and genome assemblies of each strain have been deposited in EBI/EMBL under accession numbers: ERR987451 and FIZQ01000001-34 for RC7, ERR987453 and FIZO01000001-24 for RC11, ERR987452 and FIZP01000001-42 for RC20, respectively.

First, the taxonomic position of all strains was determined using full length16S rRNA gene sequences comparisons. The 16S rRNA gene sequences of strains RC7, RC11 and RC20 were extracted from their whole genome sequences and obtained from public databases for other *Campylobacter* type strains. Sequence alignment and neighbor-joining tree construction was performed with MEGA6 [220], using 1,000 repetitions to determine bootstrap values. The 16S rRNA gene sequence similarity between the three strains was 100%, while the similarity with respect to the closest species *C. mucosalis* (ATCC 43264) was 94%. The strains formed a distinct clade between *C. mucosalis*, *C. concisus* and *C. curvus* (Fig. 7.1). Additionally, the 16S rRNA gene sequences of RC7, RC11 and RC20 were analyzed using BLAST [79] against the whole non-redundant (nr) NCBI database in order to evaluate the presence of similar strains in previous *Campylobacter* surveys by other authors. This resulted in a single top hit with 99% of identity to *Campylobacter* sp. 11S02629-5 (accession number KJ081202), isolated from the feces of *Testudo graeca*. This finding demonstrates the presence of very close strains in a distinct *Testudo* species from an independent study.

For improved resolution in species delimitation two alternative approaches were assayed. First, as the *atpA* gene has been shown as a good marker to increase intraspecies resolution within the Campylobacteraceae family [**miller2014**], we extracted the *atpA* nucleotide sequence from RC7, RC11 and RC20 and from public *Campylobacter* genomes to build a neighbor-joining tree following the same methodology as for the 16S rRNA gene. Second, from the same set of genomes

we extracted a set of 40 universal proteins of prokaryotes were extracted from RC7, RC11, RC20 and public *Campylobacter* genomes using FetchMG [221]. Then, a Maximum Likelihood tree from the alignment of concatenated amino acid sequences was built using the JTT substitution model in RAxML [227], with 1,000 bootstrap repetitions. Again, the three strains formed a well-supported clade distinct from the rest, however this analysis revealed slightly different relationships with sister clades (Fig. 7.2).

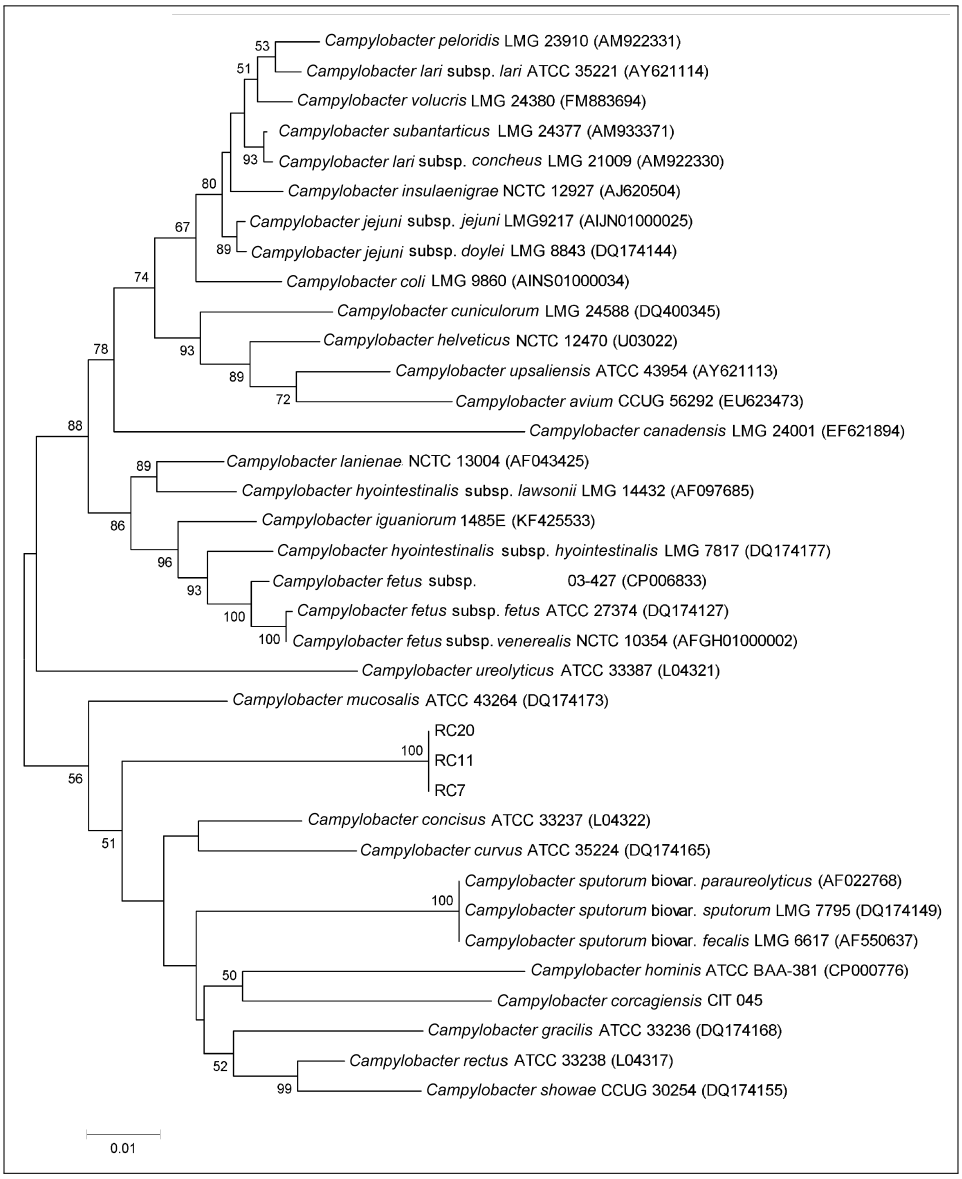


Figure 7.1: 16S phylogeny. Neighbor-joining phylogenetic tree based on 16S rRNA sequences from *Campylobacter* type strains. Bootstrap values calculated over 1,000 replications are indicated at internal nodes.

The average nucleotide identity (ANI) was calculated for species delimitation, as it can be used as an alternative for DNA-DNA hybridization (DDH) [222], where DDH species threshold of 70% corresponds to 95% ANI [253]. This analysis was performed between RC7, RC11, RC20 and all *Campylobacter* species with available genomes by implementing the original algorithm described in Konstantinidis & Tiedje (2005) inside an in-house R function that uses BLAST [79]. Strains RC7, RC11 and RC20 were extremely similar showing ANI values over 99%, while the ANI between these strains and the rest of *Campylobacter* species ranged between 78% and 80%, always far below the 95% species threshold (Tab. 7.1).

Table 7.1: Average nucleotide identity (ANI) values based on reciprocal BLASTN for *C. geochelonis* sp. nov. and most closely related *Campylobacter species*.

		1	2	3	4	5	6	7	8	9
1	RC7	100	**99**	**99**	78	79	78	77	79	78
2	RC11	**99**	100	**99**	77	79	78	77	79	78
3	RC20	**99**	**99**	100	78	79	78	78	80	78
4	*C. mucosalis* DSM 21682	78	77	78	100	77	77	78	79	77
5	*C. corcagiensis* CIT 045	79	79	79	77	100	81	78	80	79
6	*C. ureolyticus* CIT 007	78	78	78	77	81	100	81	83	78
7	*C. gracilis* RM3265	77	77	78	78	78	81	100	84	83
8	*C. hominis* ATCC BAA-381	79	79	80	79	80	83	84	100	80
9	*C. sputorum* 08/209	78	78	78	77	79	78	83	80	100

The average amino acid identity (AAI) calculations supported the ANI values. The AAI was calculated using the predicted proteomes from RC7, RC11 and RC20 which were compared with all available *Campylobacter* proteomes (predicted from public genomes) using pairwise BLASTP alignments. For each pair of taxa, identity among proteins conserved across all analyzed genomes (core proteome) was used as a measure of overall genetic relatedness. An AAI range from 65-70% was observed between the common proteome of the three strains and each of the remaining species, with *C. ureolyticus* and *C. corcagiensis* as the closest taxa (70%). This is in accordance to the phylogenetic position of RC7, RC11 and RC20 based on the tree built with 40 universal proteins for prokaryotes (Fig. 7.2). The AAI calculated between the three strains was >96% on average (Tab. 7.2). These results indicate that strains RC7, RC11 and RC20 are closely related at protein

level and also form a distinct cluster that diverges from their most related *Campylobacter* species. The genomic G+C content was calculated from the whole genome sequence using R package seqinr [**charif2005**]. Strains RC7, RC11 and RC20 showed a G+C content of 33.68%, 33.69% and 33.57%, respectively, which falls within the range reported for the genus *Campylobacter* (29-47%) [245].

Phenotypic characterization of the three *Campylobacter* strains was performed using standard methods according to the procedures described by On & Holmes [142, 254], Ursing *et al.* [255], and On *et al.* [256]. Results of phenotypic testing of the strains compared to other closely related *Campylobacter* taxa are summarized in Tab. 7.3. All strains showed oxidase and catalase activity, hydrolyzed hippurate and reduced nitrates. No urease, alkaline phosphatase or indoxyl acetate hydrolysis activity were detected. Strains did not produce haemolysis on blood agar (TSA + 5% sheep blood) and H_2S on TSI agar, but a weak production of H_2S on SIM was observed after 72 h incubation. All strains grew well on TSA agar with 5% sheep blood at 25 °C and 37 °C under microaerobic conditions and at 37 °C under anaerobic conditions, whereas no growth was observed at 18-22 and 42 °C under microaerobic conditions and at 37 °C under aerobic conditions. Additionally, they grew well on CCDA, Karmali and CAT agar, and in presence of 1% glycine and 1.5% NaCl. Strains did not grow on Muller-Hinton and MacConkey agar nor in presence of 3.5% NaCl. The addition of 0.1% selenite and 0.04% 2,3,5-triphenyltetrazolium chloride to the medium inhibited growth. All strains were susceptible to cephalotin (30 μg) and resistant to nalidixic acid (30 μg). The three strains lacked the S-layer coding genes, based on BLAST searches against known *C. fetus* sap genes. Together with genotypic characterization, the biochemical properties differentiated the tortoises' isolates from other *Campylobacter* species.

Table 7.2: Average amino acid identity (AAI) values based on reciprocal BLASTP for *C. geochelonis* sp. nov. and most closely related *Campylobacter* species.

		1	2	3	4	5	6	7	8	9
1	RC7	100	**97**	**96**	65	70	70	65	68	68
2	RC11	**97**	100	**96**	65	70	70	65	68	68
3	RC20	**96**	**96**	100	65	70	70	66	68	68
4	*C. mucosalis* DSM 21682	65	65	65	100	65	64	62	62	65
5	*C. corcagiensis* CIT 045	70	70	70	65	100	72	61	65	67
6	*C. ureolyticus* CIT 007	70	70	70	64	72	100	60	66	67
7	*C. gracilis* RM3265	65	65	66	62	61	60	100	65	60
8	*C. hominis* ATCC BAA-381	68	68	68	62	65	66	65	100	66
9	*C. sputorum* 08/209	68	68	68	65	67	67	60	66	100

The addition of 0.1% selenite and 0.04% 2,3,5-triphenyltetrazolium chloride to the medium inhibited growth. All strains were susceptible to cephalotin (30 μg) and resistant to nalidixic acid (30 μg). The three strains lacked the S-layer coding genes, based on BLAST searches against known *C. fetus* sap genes. Together with genotypic characterization, the biochemical properties differentiated the tortoises' isolates from other *Campylobacter* species.

In conclusion, results of this genotypic and phenotypic taxonomic study strongly support that the three *Campylobacter* strains isolated from western Hermann's tortoises (*Testudo hermanni hermanni*) in Northern Italy represent a novel species of the *Campylobacter* genus. The evidence on analysis of 16S rRNA genes, 40 universal proteins, ANI, AAI, growth and biochemical properties is coherent to establish RC7, RC11 and RC20 as a new species distinct from other currently described *Campylobacter* species. The description of an additional *Campylobacter* spp. from tortoises points out the importance of these animals as unique *Campylobacter* reservoirs. The strains were isolated from fecal samples of apparently healthy individuals and the host range and pathogenic potential is currently unknown, however a handful of potential virulence factors were identified in the genomes of the three strains, including genes for cytolethal distending toxin, type IV secretion systems, fibronectin-binding proteins or invasion antigen B. Considering this and that genomic signatures associated to pathogenicity and niche preferences have been identified in other *Campylobacter* species [257], a more comprehensive analysis including reptilian-derived campylobacters is imperative to uncover host-associated evo-

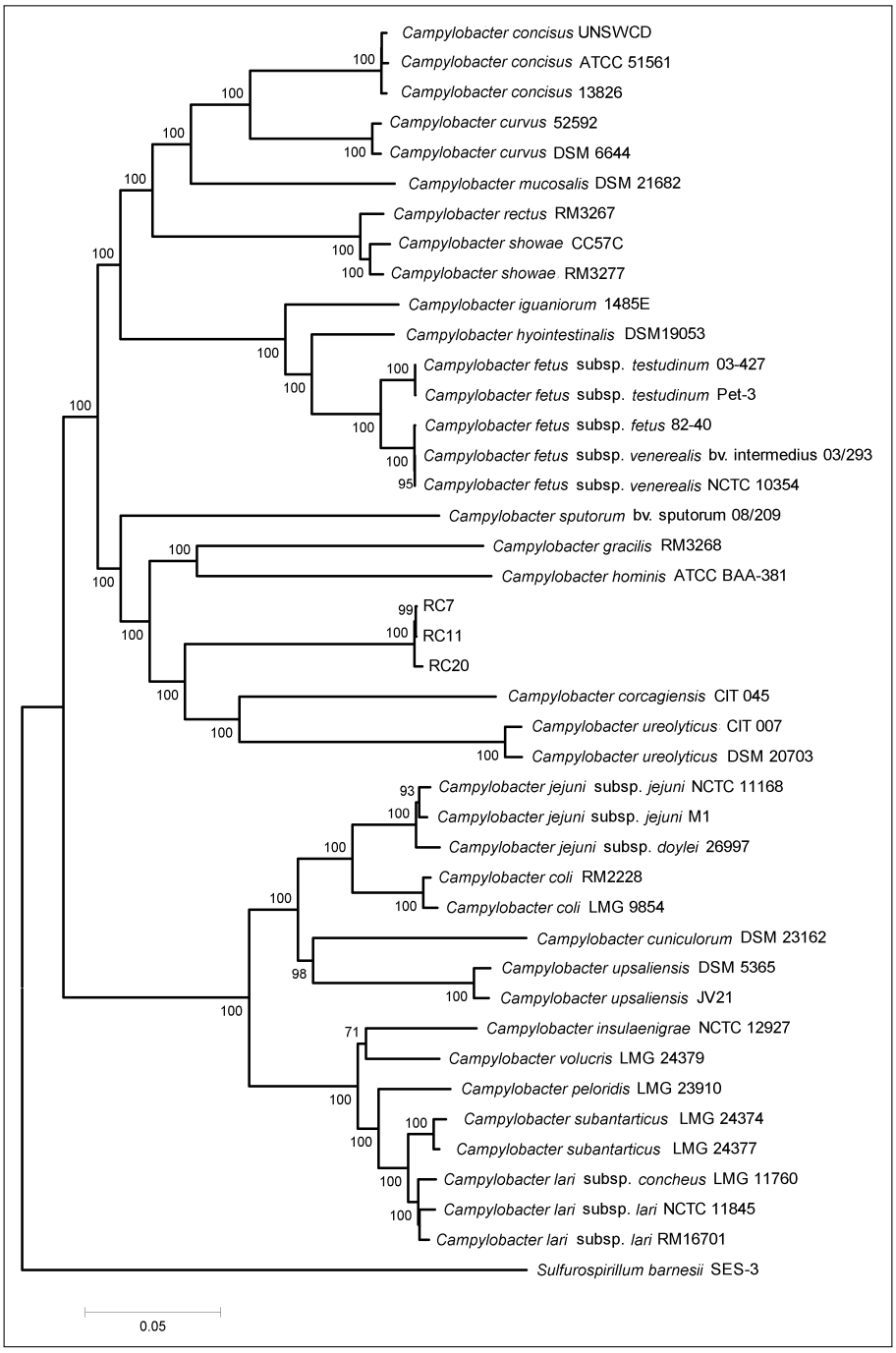


Figure 7.2: Universal proteins phylogeny. Maximum likelihood phylogenetic tree based on 40 universal proteins for prokaryotes extracted from available *Campylobacter* genomes. Bootstrap values calculated over 1000 repetitions are indicated at internal nodes. The tree was rooted with *Sulfurospirillum barnesii*.

lutionary trends in these hosts. The proposed name for this species is *Campylobacter geochelonis* sp. nov., with RC20^T (=DSM *under certification of deposit*^{T} =LMG *under certification of deposit*^{T}) as the type strain.

Table 7.3: Taxa: 1 = *C. geochelonis* sp. nov. (n = 3); 2, *C. corcagiensis*; 3, *C. gracilis*; 4, *C. hominis*; 5, *C. mucosalis*; 6, *C. sputorum*; 7, *C. ureolyticus*. Data for reference taxa were taken from the original descriptions. +, 90-100%; (+), 75-89%; V, 26-74%; (-), 11-25%; -, 0-10%; NA, not available; *, test results differ between *C. sputorum* biovar sputorum (catalase and urease negative), paraureolyticus (catalase negative, urease positive), and fecalis (catalase positive, urease negative).

	1	2	3	4	5	6	7
Oxidase	+	+	-	+	+	+	+
Catalase	+	+	V	-	-	V*	V
Urease	-	+	-	-	-	V*	-
Alkaline phosphatase	-	+	-	-	(+)	-	-
Reduction of:							
Nitrate	+	(+)	(+)	-	(-)	(+)	+
Selenite	-	NA	-	-	-	+	-
TTC	-	-	-	-	-	-	-
Hydrolysis of:							
Hippurate	+	-	-	-	-	-	-
Indoxyl acetate	-	V	V	-	-	-	V
Grow at/in/on:							
18-22 °C (microaerobic)	-	NA	-	NA	-	-	-
25 °C (microaerobic)	+	NA	-	-	-	-	-
37 °C (microaerobic)	+	+	-	+	+	+	+
42 °C (microaerobic)	-	+	V	(-)	+	+	V
37 °C (anaerobic)	+	+	+	+	+	+	V
37 °C (aerobic)	-	-	-	-	-	-	-
CCDA	+	NA	V	NA	+	(+)	+
MacConkey	-	-	(+)	-	(+)	V	V
Glycine (1%)	+	+	+	+	V	+	+
NaCl (1.5%)	+	+	+	NA	-	+	+
NaCl (3.5%)	-	NA	NA	NA	-	NA	+
Resistance to:							
Nalidixic acid (30 μg)	+	+	V	V	(+)	(+)	-
Cephalotin (30 μg)	-	NA	-	-	-	-	NA
H_2S production (TSI)	-	+	-	-	+	+	-
α-Haemolysis	-	-	-	-	-	+	V
H_2 requirement	-	-	+	+	+	-	+
S-layer presence	-	NA	-	-	-	-	-
DNA G+C content (mol %)	33.6	31.9	44-46	32.5	36-38	29-33	28-30

7.3 Description of *Campylobacter geochelonis* sp. nov.

Campylobacter geochelonis [geo.che.lo.ni's. Gr. pref. gèo-, earth; Gr. n. masc. sing., chelòne, turtle; Gr. masc. adj., geochelonis, pertaining to terrestrial tortoise, including *Testudo hermanni hermanni* the reptile from which the bacterium was isolated].

Gram negative, most cells are straight rods, some are slightly curved, 1 to 2 um long. Motile. After incubation on blood agar (TSA + 5% sheep blood) at 37°C for 48 h under microaerobic conditions, colonies are grey and translucent in colour, tiny (0.1 to 1 mm in diameter), smooth, round and not haemolytic. In old cultures, colonies may grow up to 3 mm in diameter. Growth is observed on blood agar at 37°C under microaerobic and anaerobic (slightly weaker growth), but not aerobic conditions. Strains do not require H_2 to grow, but are able to grow in its presence. Strains show growth at 25°C under microaerobic conditions, not at 18-22°C and 42°C. All strains are oxidase and catalase-positive, urease-negative. Indoxyl acetate is not hydrolysed, hippurate is hydrolysed by all strains. All strains reduce nitrates, selenite and TTC. Hydrogen sulphide is not produced on TSI agar, but a weak production of H_2S on SIM agar following 72 h incubation is observed. Growth is observed on blood agar containing 1.5% NaCl and 1% glycine, no growth in the presence of 3.5% NaCl. Strains grow well on CCDA, Karmali and CAT agar, not on Muller Hinton and MacConkey agar. All strains are resistant to nalidixic acid (30 ug) and susceptible to cephalotin (30 ug). Pathogenicity is unknown. Strains were isolated from healthy western Hermann's tortoises (*Testudo hermanni hermanni*) in Northern Italy.

The RC20 (= DSMZ *under certification of deposit*^{T} = NCTC *under certification of deposit*^{T}) has been designated as the type species and was isolated from a western Hermann's tortoise in Northern Italy in 2011. Strains RC7 (= DSMZ *under certification of deposit*^{T} = NCTC *under certification of deposit*^{T}) and RC11 (= DSMZ *under certification of deposit*^{T} = NCTC *under certification of deposit*^{T}) belong to the same species and were isolated from the western Hermann's tortoise. The pathogenicity potential of these strains is unknown.

7.4 Acknowledgments

We thank to Dr. Martina Giacomelli for performing the initial screening of reptile samples.

7.5 References

[79] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **1990**, *215*, 403–410.

[86] G. Iraola, M. Hernandez, L. Calleros, F. Paolicchi, S. Silveyra, A. Velilla, L. Carretto, E. Rodriguez, R. Perez, *J. Vet. Sci.* **2012**, *13*, 371–376.

[138] M Veron, R Chatelain, *International Journal of Systematic and Evolutionary Microbiology* **1973**, *23*, 122–134.

[142] S. On, B Holmes, *Journal of clinical microbiology* **1992**, *30*, 746–749.

[145] S. M. Man, *Nat Rev Gastroenterol Hepatol* **2011**, *8*, 669–685.

[147] D. Linton, R. J. Owen, J. Stanley, *Res. Microbiol.* **1996**, *147*, 707–718.

[149] D. R. Zerbino, E. Birney, *Genome Res.* **2008**, *18*, 821–829.

[202] C. Fitzgerald, Z. C. Tu, M. Patrick, T. Stiles, A. J. Lawson, M. Santovenia, M. J. Gilbert, M. van Bergen, K. Joyce, J. Pruckler, S. Stroika, B. Duim, W. G. Miller, V. Loparev, J. C. Sinnige, P. I. Fields, R. V. Tauxe, M. J. Blaser, J. A. Wagenaar, *Int. J. Syst. Evol. Microbiol.* **2014**, *64*, 2944–2948.

[203] S. Hum, K. Quinn, J. Brunner, S. L. On, *Aust. Vet. J.* **1997**, *75*, 827–831.

[219] T. Seemann, *Bioinformatics* **2014**, *30*, 2068–2069.

[220] K. Tamura, G. Stecher, D. Peterson, A. Filipski, S. Kumar, *Mol. Biol. Evol.* **2013**, *30*, 2725–2729.

[221] S. Sunagawa, D. R. Mende, G. Zeller, F. Izquierdo-Carrasco, S. A. Berger, J. R. Kultima, L. P. Coelho, M. Arumugam, J. Tap, H. B. Nielsen, S. Rasmussen, S. Brunak, O. Pedersen, F. Guarner, W. M. de Vos, J. Wang, J. Li, J. Dore, S. D. Ehrlich, A. Stamatakis, P. Bork, *Nat. Methods* **2013**, *10*, 1196–1199.

[222] K. T. Konstantinidis, J. M. Tiedje, *J. Bacteriol.* **2005**, *187*, 6258–6264.

[227] A. Stamatakis, *Bioinformatics* **2014**, *30*, 1312–1313.

[244] P. Vandamme, F. E. Dewhirst, B. J. Paster, S. L. On, *BergeyâĂŹs Manual® of Systematic Bacteriology* **2005**, 1161–1165.

[245] L. Debruyne, D. Gevers, P. Vandamme, **2008**.

[246] N. O. Kaakoush, N. Castaño-Rodríguez, H. M. Mitchell, S. M. Man, *Clinical microbiology reviews* **2015**, *28*, 687–720.

[247] M. J. Gilbert, M. Kik, W. G. Miller, B. Duim, J. A. Wagenaar, *International journal of systematic and evolutionary microbiology* **2015**, *65*, 975–982.

[248] M Giacomelli, A Piccirillo, *Veterinary Record* **2014**, vetrec–2013.

[249] M. Giacomelli, C. Andrighetto, A. Lombardi, M. Martini, A. Piccirillo, *Avian diseases* **2012**, *56*, 693–700.

[250] W. Yamazaki-Matsune, M. Taguchi, K. Seto, R. Kawahara, K. Kawatsu, Y. Kumeda, M. Kitazato, M. Nukina, N. Misawa, T. Tsukamoto, *Journal of medical microbiology* **2007**, *56*, 1467–1473.

[251] M. Maiwald, *Molecular microbiology: diagnostic principles and practice. 2nd ed. Washington DC: American Society of Microbiology* **2004**, 379–90.

[252] K Willoughby, P. Nettleton, M Quirie, M. Maley, G Foster, M Toszeghy, D. Newell, *Journal of applied microbiology* **2005**, *99*, 758–766.

[253] J. Goris, K. T. Konstantinidis, J. A. Klappenbach, T. Coenye, P. Vandamme, J. M. Tiedje, *International journal of systematic and evolutionary microbiology* **2007**, *57*, 81–91.

[254] S. On, B Holmes, *Journal of clinical microbiology* **1991**, *29*, 923–926.

[255] J. B. Ursing, H. Lior, R. J. Owen, *International Journal of Systematic and Evolutionary Microbiology* **1994**, *44*, 842–845.

[256] S. On, B Bloch, B Holmes, B Hoste, P Vandamme, *International Journal of Bacteriology* **1996**, *45*, 767–774.

[257] G. Iraola, R. Pérez, H. Naya, F. Paolicchi, E. Pastor, S. Valenzuela, L. Calleros, A. Velilla, M. Hernández, C. Morsella, *Genome biology and evolution* **2014**, evu195.

Conclusión del Capítulo 7

En este trabajo se describe una nueva especie del género *Campylobacter*, denominada *Campylobacter geochelonis*. Las cepas fueron aisladas de la tortuga de tierra *Testudo hermanni hermanni*. Inicialmente, las cepas que luego fueron reclasificadas como *C. geochelonis* fueron identificadas como *C. fetus* subsp. *fetus* debido a sus características similares desde el punto de vista bioquímico y bacteriológico.

La secuenciación del genoma completo de las cepas y su comparación con las especies descritas dentro del género, aportaron la evidencia necesaria para establecer la presencia de una especie nueva. Las diferencias presentes en los genomas luego fueron corroboradas por métodos filogenéticos y bioquímicos que sustentaron la designación de *C. geochelonis*.

Las uńicas especies del género *Campylobacter* descritas a partir de reptiles hasta el momento son *C. fetus* subsp. *testudinum* y *C. iguaniorum*. La identificación de más especies en estos hospederos, como *C. geochelonis*, evidencia la importancia de los reptiles como reservorio de especies bacterianas potencialmente relevantes en la salud humana y animal.

Parte III

Estudios genómicos en otros organismos

(Genomic studies in other organisms)

8 The transcriptome of *Leptospira biflexa* biofilms

Citation:

Iraola G, Spangenberg L, Lopes Bastos B, Graña M, Vasconcellos L, Almeida AM, Greif G, Robello C, Ristow P, Naya H*. (2016) **Transcriptome sequencing reveals wide expression reprogramming of basal and unkwnown genes in *Leptospira biflexa* biofilms**. *mSphere*. 1(2):e00042-16.

* Corresponding authors

8.1 Abstract

The genus *Leptospira* is composed by pathogenic and saprophytic spirochetes. Pathogenic *Leptospira* are the etiological agent of leptospirosis, a globally spread neglected disease. A key ecological feature of pathogenic leptospires is their ability to survive both within and outside the host. For most leptospires, the ability of persisting outside the host is associated with biofilm formation, a most important bacterial strategy to face and overcome hostile environmental conditions. The architecture and biochemistry of leptospiral biofilms are rather well understood, however, the genetic program underpinning biofilm formation remains mostly unknown. In this work, we used the saprophyte *Leptospira biflexa* as a model organism to assess over- and down-represented transcripts during biofilm state, using RNA-seq technology. Our results showed that some basal biological processes like DNA replication and cell division are down-regulated in the mature biofilm. Additionally, we identified significant expression reprogramming for genes involved in motility, sugar/lipid metabolism and iron scavenging, as well as coding for outer membrane coding genes. A careful manual annotation process allowed us to assign molecular functions to many previously uncharacterized genes that are probably involved in biofilm metabolism. We also provided evidence for the presence of small regulatory RNAs in this species. Finally, co-expression networks were reconstructed to pinpoint functionally related gene clusters that may explain how biofilm maintenance is regulated. Beyond elucidating some genetic aspects of biofilm formation, this work reveals a number of pathways whose functional dissection may impact our understanding of leptospiral biology, in particular how these organisms adapt to environmental changes.

8.2 Introduction

Leptospirosis is a neglected disease caused by infections with bacteria belonging to the genus *Leptospira*. This worldwide-distributed zoonotic disease is relevant for animal and human health, with more than 500,000 documented cases per year and particularly incident in

developing countries [258]. The genus *Leptospira* contains both saprophytic and pathogenic species differing in their capacities for surviving and colonizing different environments and hosts, ranging from soil and water to mammalian tissues during infection [259]. *Leptospira* species have been historically classified in three groups according to their pathogenic potential: "pathogens", "intermediate pathogens" and "saprophytic" [260]. The advent of genomics allowed to identify 21 species that are phylogenetically correlated with the previously referred groups. Recently, a revision of leptospiral taxonomy based on genomics proposed the following classification: Group I (previously known as "pathogens") include 9 species that comprise *L. interrogans*, *L. kirschneri* and *L. noguchii*, which cause the most severe cases of leptospirosis. Group II ("intermediate pathogens") include 5 species that predominantly cause milder cases of leptospirosis. Group III ("saprophytic") is conformed by non-pathogenic, free-living environmental leptospires like *L. biflexa* [261]. This classification is herein adopted.

The increasing availability of whole genome sequences for species belonging to the three groups has enabled the identification of genome-wide evolutionary processes involved in the transition from a non-pathogenic and free-living to a pathogenic and host-adapted lifestyle. For example, comparative genomics have revealed that *L. interrogans* (Group I) has a larger genome compared to *L. biflexa* (Group III), probably reflecting additional genetic features required for survival in both soil/water and mammalian hosts [259]. Importantly, the fact that *L. interrogans* retained the ability of surviving in the environment as a free-living organism directly impacts on the ecology and epidemiology of leptospirosis, since these organisms are capable of colonizing and multiplying inside the renal tubules of chronically infected reservoir species, disseminating in the urine and contaminating soil and water. Humans and other mammals are then infected by direct contact with animal fluids or contaminated water [260].

As stated before, survival outside the host is a key aspect of leptospiral ecology, hence for pathogenesis. As most prokaryotes, *Leptospira* can form biofilms to survive when cells are exposed to the outside environment. These matrix-confined bacterial populations protect single cells from adverse conditions, favoring persistence and transmission of infectious diseases [262]. The transition between planktonic and

biofilm phenotypes occurs as a response to various environmental signals. It involves producing and assembling components of an extracellular matrix, cell migration, adhesion and aggregation, among other processes, which are regulated by the expression of specific genes. In this sense, the consolidation of whole RNA sequencing (RNA-seq) as the gold standard method for evidencing transcription reprogramming through biological conditions [263], has enabled the study of differential gene expression associated to biofilm formation in many microorganisms [264–266]. Nonetheless, and despite biofilm formation has been described *in vitro* for pathogenic and saprophytic leptospires [267], and also observed *in vivo* [268], a genome-wide transcriptomic analysis is still lacking for *Leptospira* species in the context of biofilm formation.

From an ecological point of view, leptospiral pathogenesis can be linked with biofilms particularly in species that can complete a life cycle within and outside the host, so elucidating the genetic basis of biofilm formation can provide useful tools for genetic manipulation, drug design and vaccine development, which should directly impact on disease handling and could substantially improve the design of preventive schemes. In this work, we selected *L. biflexa* serovar Patoc strain Patoc 1 (Paris) as model organism to compare the global gene expression profile between biofilm grown on abiotic surfaces and planktonic cells, using RNA-seq. Our results indicate that around 99% of genes automatically annotated in *L. biflexa* genome are being transcribed. Biofilm growth requires the extensive reprogramming of transcription patterns along the three replicons of *L. biflexa*, and involves many regulatory networks like c-di-GMP signaling, anti-anti-sigma factors and canonical two-component systems that control basal functions, like DNA metabolism and replication, as well as more specific functions like cell motility or lipid and sugar metabolisms.

8.3 Methods

8.3.1 *Leptospira biflexa* cultures and biofilm experiments. *Leptospira biflexa* serovar Patoc strain Patoc1 (Institut Pasteur Paris) was gifted from Centro de Pesquisas GonÐalo Moniz (CPqGM), Fundação Oswaldo Cruz (Fiocruz), Bahia, Brazil. Bacteria were cultured in

Ellinghausen, McCullough, Johnson & Harris (EMJH) liquid medium (Difco, USA) at 29°C, without shaking. *L. biflexa* was replicated without shaking ten times in liquid EMJH before performing biofilm experiments.

Biofilms were grown in borosilicate glass tubes (16 mm X 100 mm) containing 5 mL liquid EMJH. A starting culture in mid-exponential growth phase (~107 leptospires/mL, after 48 h incubation) was expanded to 30 tubes, each containing 5 mL liquid EMJH (1:10 v/v), making six biological replicates of five tubes each. Biofilms were harvested at two time points: 1) after 48 h of incubation, when biofilms are considered to be in a mature stage, and a dense halo is visible attached to the wall of glass tubes at the air-liquid interface (hereinafter referred to mature biofilm), and 2) after 120 h of incubation, in a late culture stage, when biofilms are detaching (hereinafter stated as late biofilm). Biofilms were visually inspected using dark-field microscopy by removing the biofilm mass from the tube wall in order to check for cell motility, aggregation/detachment and biofilm mass integrity. At 48 h and 120 h, three biological replicates were randomly chosen. Liquid EMJH was discarded and the biofilms were rinsed with 6 mL cold liquid EMJH to remove unattached bacteria. To each glass tube, 400 μl RNA Protect Reagent (Qiagen, USA) was added and biofilms were scraped using stainless steel sterile spatulas. The unavoidable destruction of the biofilm heterogeneity during sample preparation prevents the study of gene expression patterns across different populations within the biofilm, hence the results obtained will reflect an average expression pattern of the whole biofilm. Planktonic cells were cultured in polypropylene tubes. A starting culture with 48 h incubation (~107 leptospires/mL) was replicated to six polypropylene tubes containing 10 mL liquid EMJH each (1:10 v/v). At 48 h and 120 h, three tubes, representing three biological replicates, were randomly selected. From each tube, 1 mL of planktonic culture was transferred to another plastic tube containing 2 mL of RNA Protect Reagent (Qiagen, USA).

8.3.2 RNA purification and sequencing. Total RNA for each biological condition and replicate was isolated using RNeasy Protect Bacteria Mini Kit (Qiagen, USA), according to manufacturer's protocol. For the planktonic condition, 1 mL of liquid culture medium was used as

starting material. For the biofilm condition, the biofilm mass was mechanically removed from the glass tube and homogenized in 1 mL of PBS. Ribo-Zero Magnetic Kit (Bacteria) (Epicentre, USA) was used to deplete ribosomal RNA from 1 μg of total RNA. Obtained ribosomal-depleted RNA was quantified with Qubit^{TM} RNA HS assay kit (Invitrogen, USA). ScriptSeqTM v2 RNA-Seq Library Preparation Kit (Epicentre, USA) was used from 50 ng of ribosomal-depleted RNA. Index primer were added to each library to allow sequence multiplexing. After 12 PCR cycles, the final library was purified with AMPure XP (Benchman, USA) and quantified with Qubit^{TM} dsDNA HS Assay Kit (Invitrogen, USA). Quality and length of the library was assessed with Agilent High Sensitivity DNA Kit (Agilent, USA) using the 2100 Bioanalyzer (Agilent, USA). Sequencing was performed on an Illumina Genome Analyzer II X platform at the Institut Pasteur Montevideo and generated 45,365,550 single-end reads (72 cycles). Data was deposited in the Sequence Read Archive (SRA) database, accession numbers are listed in Supp. Tab. 8.1.

8.3.3 Detection of differentially expressed genes. All statistical analyses were implemented in R. Read alignment and counting was performed using the Rsubread package [269]. Read duplicates were kept as for most samples, at the reached coverage, more than one "real" duplicate is expected at each starting position. The minimum, mean and maximum reads number per sample was 2.631.490, 3.780.463 and 7.500.998, respectively; the proportion of mapped reads was greater than 98.6% for all samples. Differential expression analysis was carried out with the edgeR package [270]. After previous analysis and visual inspection, 3 samples (BC48, PA48, PC120) were discarded because they showed discordant expression patterns when considering time and source, probably due to problems inherent to cells manipulation. Genes with less than one count per million (CPM < 1) in any of the samples were also discarded (15 genes). A simple factorial model with two factors, Time (48 and 120 h) and Mode (Planktonic and Biofilm), was fitted; False Discovery Rate (FDR) < 1e-2 was considered as threshold for differentially expressed genes.

Small RNA genes (sRNAs) were predicted with RNAspace [271], that combines the results of several de novo prediction tools for RNAs.

Parameters were set as default and all predicted RNAs were kept at first. Genes with CPM > 1 or with at least 2 reads in at least 2 samples were defined as transcriptionally active. The identification of differentially expressed sRNAs was performed as described in the previous paragraph.

8.3.4 Functional annotation and co-expression analyses. When interesting genes were annotated as hypothetical proteins in the current version of *L. biflexa* strain Patoc I genome, additional efforts were made to predict molecular functions. In first place, runs of Blastp and CDD-search against the nr database (NCBI) were used to identify annotated bacterial homologs. Additionally, over the remaining set of proteins with unknown function, structural annotations were attempted using the HH-suite package and database [272]. Briefly, for each query protein, a profile Hidden Markov Model (HMM) was built using HHblits [273], with three round searches over a non-redundant HMM database. Next, the resulting HMM was used as a query over the HMM database for the Protein Data Bank culled at 70% sequence identity (PDB70) provided by the authors, using the HHsearch program [272]. Results were manually inspected and, when possible, a structural/functional feature was assigned to the query protein.

A simple analysis of co-expression networks was performed considering the correlation matrices of gene expression (CPM) across samples. For selected genes, a correlation coefficient value greater than 0.96 was arbitrarily set as threshold for gene clustering. Upon visual inspection and analysis of cluster contents, only positive correlations were graphically represented using igraph R package [274].

8.3.5 Confirmation of differentially expressed genes by RT-PCR. Twenty one genes were selected for testing their expression levels using a Real-Time PCR protocol for relative transcript quantification. This experiment was preformed using RNA purified from an independent experiment, different than the one used for RNA-seq to check the robustness and reproducibility of the results. For all samples, 100 ng of total RNA was used to synthesize first-strand cDNA with reverse transcriptase SuperScript II (Invitrogen, USA) and oligo-dT. The cDNA synthesis was performed at 42°C for 50 min after heat inactivation at 70°C

for 10 min. The primer sequences designed for selected genes are listed in Supp. Tab. 8.2. PCR was performed using 1× KAPA SYBR®FAST qPCR Kit Master Mix (Kappa, USA) on an Illumina EcoTM machine (Illumina, USA). For all genes, cycling conditions were as follows: 2 min at 95°C, and 40 cycles of 10 s at 95°C, plus 30 s at 60°C. The Eco study software (Illumina, USA) was used to calculate ΔΔCt relative expression values for all the genes studied. For endogenous normalization of expression levels we selected a set of 6 genes (Supp. Table 8.3) that showed the lowest count variation among samples in the RNA-seq experiment. As differences in the performance of all genes as normalizers were not significant, we selected one of them (*LEPBI_I2771*) for presenting RT-PCR results.

8.4 Results and Discussion

8.4.1 Transcriptomic overview of *L. biflexa*. The whole transcriptome was sequenced for 12 cultures of *L. biflexa* Patoc strain Patoc I harvested at 48 h and 120 h in both biofilm and planktonic culture conditions, using biological triplicates. The average reads yield per sample was ~4 million, indicating a sufficient amount of data for performing differential expression analyses [275, 276]. Out of the total number of reads sequenced per sample, ~99% mapped against the reference *L. biflexa* Patoc strain Patoc I genome (Supp. Tab. 8.4). The hierarchical clustering of samples using normalized reads counts was consistent with harvesting time and culture condition, upon removal of three discordant samples, not included in subsequent analyses. For differential gene expression analyses, all possible comparisons of time (mature or late) and culture conditions (planktonic or biofilm) were carried out using the 9 consistently clustering samples (Supp. Fig. 8.2). The number of down- and up-regulated genes (FDR $<1\times10^{-2}$) for each comparison are shown in Tab. 8.1. The most relevant information for identifying functional changes in gene expression came from comparing mature biofilm vs. mature planktonic cells.

Table 8.1: Number of differentially expressed genes detected in each comparison at FDR $<1\times10^{-2}$.

Comparison	Up	Down	Total
BvsP_48	121	198	319
BvsP_120	184	117	301
B_48vs120	151	172	323
P_48vs120	184	240	424

The reference genome of *L. biflexa* Patoc strain Patoc I encodes a total of 3,771 predicted genes distributed within three replicons: chromosome I (CI), chromosome II (CII) and a 74-kb plasmid (P74) with chromosome-like features [259]. Transcriptional activity was detected in 3,762 genes in at least one sample, indicating that the vast majority (99%) of predicted genes of *L. biflexa* were transcriptionally active. Further analysis of 9 annotated genes that remained silent in all samples evidenced the presence of small hypothetical proteins and RNA-coding genes. Most notoriously, two pairs of the MerR/MerT system were identified among these silent genes. This system belongs to the mer operon, involved in the resistance to high concentrations of mercury ions and organic compounds containing this metal [277]. These are the unique two copies of *merR* and *merT* genes in the genome of *L. biflexa*, interspaced by ~36 kb in CI. Both gene pairs show the same arrangement and are surrounded by a number of hypothetical genes. However, one pair is closer to putative plasmid-like genes, suggesting horizontal acquisition and/or gene duplication. Dissecting why this system remains totally silent in *L. biflexa* will require further investigation.

8.4.2 Expression through replicons. The number of differentially expressed genes varied when considering mature (48 h) or late (120 h) biofilms and also when considering gene location (CI, CII or P74). For instance, in mature biofilms, up-regulated genes only came from CI and P74 (Fig. 8.1A), while down-regulated genes were exclusively found in CII (29% of encoded genes in this chromosome) (Fig. 8.1B). These results suggest that replicons in *L. biflexa* fulfill different tasks during biofilm formation, and that up- and down-regulation is appreciably compartmentalized throughout this growing condition. This

notion is reinforced when examining late biofilms (120 h). During this stage both up- and down-regulated genes were almost equally distributed between CI and CII (Fig. 8.1B), however around 68% of genes present in P74 were up-regulated. It is not clear whether P74 behaves as a chromosome or as an extra-chromosomal element, even if some essential survival genes (like *recBCD*) are located in that replicon. These genes are found in CI in other pathogenic species like *L. interrogans* and altering their sequences has been linked to lower viability in other bacterial species, suggesting that P74 is essential for survival of *L. biflexa* [259]. Our findings support this hypothesis, considering the pervasive up-regulation of most genes coded in P74, and also suggest a previously unknown role of this replicon in the late stages of biofilm, that are featured by cells recycling, disaggregation and death.

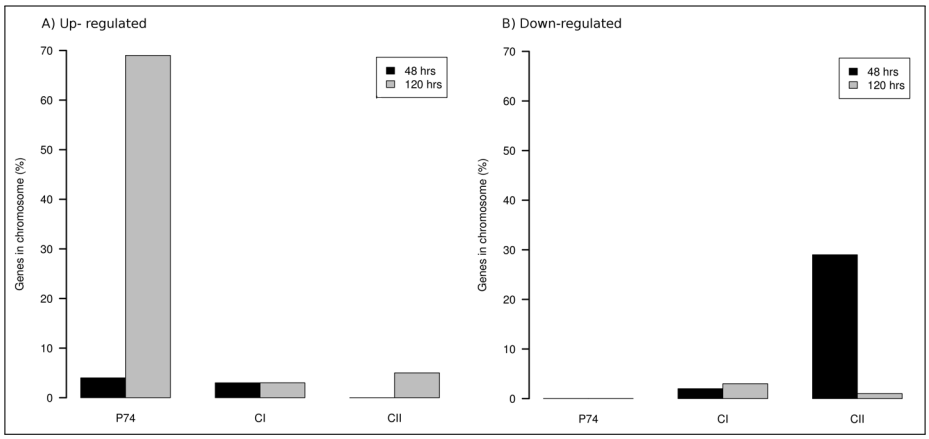


Figure 8.1: Number of genes (percentage) in each replicon with differential expression. The barplots show the percentage of each *L. biflexa* replicon (measured as number of differentially expressed genes over total number of genes in the replicon) that were A) up-regulated and B) down-regulated. In both cases black is 48 h and 120 h.

8.4.3 Replication and cell growth. The capacity of persisting in resource-limiting conditions (like environmental water in the case of *Leptospira*) is a major advantage conferred by biofilms. This aptitude is based on an altruistic behavior that relies on maximizing the biomass formed per amount of resources used [278], meaning that single cells can reduce their growth rate and resource consumption in the benefit of the whole population (biofilm). In the context of this hypoth-

esis, we found that key genes involved in DNA replication and cell division were differentially expressed. In particular, the gene coding for the chromosomal replication initiator protein DnaA (*LEPBI_I0001*) was down-regulated during mature biofilm, as well as other genes coding for proteins implied in replication, like DNA polymerase III subunits (*LEPBI_I0012*, *LEPBI_I3461* and *LEPBI_I3479*), chromosome partitioning protein ParB (*LEPBI_I3473* and *LEPBI_II0026*), replication proteins GidA and GidB (*LEPBI_I3477* and *LEPBI_3475*), DNA replication and repair protein RecF (*LEPBI_I0003*) and DNA gyrase GyrB1 (*LEPBI_I0005*). Additionally, we found up-regulated one putative gene for the virulence-associated protein of unknown function VagC (*LEPBI_I2249*) during mature biofilm. This gene presented homology to *mazE*, belonging to the MazF-MazE toxin/antitoxin system, and was also placed next to a hypothetical protein-coding gene (*LEPBI_I2248*) with homology to MazF. This system is involved in cell growth regulation and programmed cell death during resources shortage in Escherichia coli [279] and, despite the cognate MazF homolog was not transcriptionally altered, just altering the levels of MazE is enough to regulate cell growth [280]. Moreover, we found three additional down-regulated genes in mature and late biofilms that code for HepA, Fis and a pyrrolo-quinoline quinone (*LEPBI_I3440*, *LEPBI_I0011* and *LEPBI_I3348*, respectively) that have been associated to cell growth control in other bacteria. In particular, HepA and Fis have been identified as over-expressed genes during fast-growing or exponential growth phase [281, 282], while here we found them down-regulated, in accordance with the notion of low replication and cell growth in *L. biflexa* biofilms.

8.4.4 Lack of translational motility. Motility is a central paradigm in bacterial physiology. In *Leptospira* this mechanism is mainly controlled by two periplasmic flagella, whose opposite rotation provokes a topological change in both cell poles (spiral-hook configuration) that allows translational displacement by a corkscrew movement [283]. Switching from motile to non-motile forms depends on the fine interaction between the flagellar apparatus and chemotactic systems. Using dark-field microscopy we observed null translational motility in the vast majority of cells in mature and late biofilms. This

observation led us to hypothesize that genes involved in determining the spiral-hook configuration needed for translational movement were altered in the biofilm condition. We found two genes encoding *pilZ* homologs (*LEPBI_I0008* and *LEPBI_II0088*) consistently down-regulated during mature biofilm. The interaction of PilZ proteins with the flagellar switch-complex proteins FliG and FliM induces counterclockwise motor bias that results in reversing the flagellar rotation [284]. The fact that *LEPBI_I0008* and *LEPBI_II0088* remained down-regulated supports our hypothesis that spiral-hook configuration is prevented by reducing the interaction of PilZ with FliG and FliM.

The motor switch proteins FliG and FliM also have other interactors that affect flagellar motor bias, such as the signal transducer CheY. This protein presents four annotated paralogs in the genome of *L. biflexa* (*cheY1-4*), but only *cheY1* (*LEPBI_I0917*) showed differential expression (up-regulation) in mature biofilm in our analysis. The role of CheY in motility behavior has been studied using recombinant *E. coli* to evaluate *cheY* genes encoded by *L. interrogans* [285], where they are also highly redundant (5 paralogs). The overexpression of *cheY* genes from *L. interrogans* in *E. coli* mainly caused swarming inhibition [286]. Moreover, we found that *cheR* (*LEPBI_I1764*) was down-regulated in mature biofilm. The deletion of this gene in *L. interrogans* resulted in a swarming defective phenotype [287]. Based on these results, we can suggest that the up-regulation of *cheY1* and down-regulation of *cheR* should be contributing to the lack of translational movement observed in *L. biflexa* biofilms. In addition, leptospires are attached to one another in the biofilm and enclosed by an exopolysaccharidic matrix, what hampers translational motility once biofilm is mature [267].

When considering the structural components of the flagellar filament it was striking to find that *flaB123* (core flagellar subunits) (*LEPBI_I1589*, *LEPBI_I2133*, *LEPBI_I2132*, respectively) and *flaA1* (sheath subunit) (*LEPBI_I2335*) were up-regulated in mature biofilm, while no additional genes coding for the flagellar apparatus were differentially expressed in any comparison. Leptospires have only two periplasmic flagella, and despite FlaB is essential for its correct assembly and FlaA is required for motility and virulence in *L. interrogans* [287], it is difficult to interpret the possible role that overproduction of flagellar filament components may imply in the context of motility.

One possibility is that FlaA and FlaB have unknown pleiotropic functions for biofilm homeostasis.

8.4.5 Over-expression of genes for outer membrane proteins. Outer membrane proteins (OMPs) deserve great interest in *Leptospira* and bacteria in general because they are located on the cell surface, where the microorganism interacts with the environment, acting as adhesins, antigens, transporters or receptors [288]. We found several up-regulated OMP-coding genes in mature biofilms.

Probably, the most interesting up-regulated gene in this context was *LEPBI_Ia0817* that encodes the outer membrane porin OmpL1, which is a novel leptospiral extracellular matrix (ECM)-binding protein and plasminogen receptor [289]. This protein is expressed during infection [290] and presented synergistic immune protection with the lipoprotein LipL41 in *Leptospira kirschneri* infection in hamsters [288]. Considering this, the overexpression of the *L. biflexa* ompL1 suggests that it could also plays an important role in the establishment and maintenance of biofilm structure by providing adhesive properties. Another interesting up-regulated gene (*LEPBI_I1873*) encodes an OmpA-like protein exclusively present among Group III (91% average amino acid identity), with a distant homolog in *Leptonema illini* (43% amino acid identity). The over-expression of OmpA homologs has been identified as important for cell aggregation during biofilm formation in other bacterial species, such as *E. coli* [291] and *Acinetobacter baumannii* [266] suggesting that *LEPBI_I1873* may be also implied in biofilm aggregation in *L. biflexa* as well.

Five additional genes that code for putative surface-exposed lipoproteins were also identified as up-regulated. In particular, *LEPBI_I0009* encodes a surface exposed lipoprotein confined to Group III *Leptospira* (98% average amino acid identity), being LipL21 the closest protein encoded in pathogenic species from Group I and II (68% average amino acid identity). LipL21 is an abundant OMP detected in vivo during pathogenic *Leptospira* infection [292, 293], but absent in the saprophytic *L. biflexa*, suggesting that, in fact, *LEPBI_I0009* is a different protein restricted to saprophytes. Another similar case was *LEPBI_I1822* that encodes a conserved lipoprotein among Group III leptospires (92% average amino acid identity), being LipL31 its clos-

est protein in pathogenic species (52% average amino acid identity). Furthermore, the gene *LEPBI_I2674* that encodes the apolipoprotein N-acyltransferase LntB was up-regulated in mature biofilm. This protein is involved in lipoprotein biosynthesis and its depletion provokes mislocalization of outer membrane lipoproteins [294]. The overexpression of lntB has been also reported during biofilm formation of *Leptospirillum* [295], constituting additional evidence for the importance of lipoproteins in the development and maintenance of biofilms.

The transcriptional shift of these genes allowed us to hypothesize that molecular mechanisms of biofilm formation can have different actors in saprophytic and pathogenic leptospires. Future work on transcriptomics using pathogenic species during biofilm formation could shed light on these differences. None of the genes discussed in this section were differentially expressed in late biofilm. The main difference observed when comparing mature and late biofilms using darkfield microscopy was that late cultures presented evident signs of detachment, like less dense biofilm mass, interspersed cellular aggregates with areas devoid of cells and the presence of planktonic cells. This fact supports that over-expression of OMPs and surface exposed lipoproteins may be implied in the structural maintenance of mature biofilms by promoting cell aggregation and adhesion to abiotic or biotic surfaces.

8.4.6 Metabolism of sugars and lipids. Sugars and lipids are essential cellular building blocks, but also are the main carbon sources for energy production and storage. We found evidence that both sugars and lipids are mainly used to build biofilm matrix components, in particular exopolysaccharides (EPS) and fatty acids.

Galactose is a monosaccharide that can be used as carbon source via the Leloir pathway, composed by three main enzymes: GalK, GalT and GalE. The reduced activity of the epimerase GalE (the last step of the pathway) leads to accumulate UDP-galactose, which is toxic for the cell. Recently, Chai *et al.* [296] demonstrated for *Bacillus subtilis* that *galE* null mutants presented compensatory mutations in the major biofilm repressor *sinR* that overcomes UPD-galactose cytotoxicity. These mutants were characterized by an increased capacity of producing EPS, a major biofilm matrix component. In *L. bi-*

flexa, *galK* (*LEPBI_I0073*) and one *galE*-like gene coding for an UDP-glucose 4-epimerase (*LEPBI_I0113*) were down-regulated in mature biofilms (Tab. 8.2). This suggests galactose is not being fully metabolized in that condition and supports the notion of UDP-galactose accumulation. Furthermore, the pioneering characterization of *L. biflexa* lipopolysaccharide (LPS) demonstrated that galactose is abundant in this macromolecule [297], which was further confirmed [298]. In addition, it is known that the first step of O-antigen biosynthesis is limited to the incorporation of UDP-NAc-glucosamine or UDP-galactose [299]. Despite *sinR* homologs have not been identified in *Leptospira*, our results suggest that galactose metabolism could play a central role in EPS production and biofilm formation using an analogous mechanism to *B. subtilis* and that maybe galactose acts as a modulator of other regulatory genes from *L. biflexa*. Also, two additional genes (*LEPBI_I0037* and *LEPBI_I2021*) related to galactose metabolism were up-regulated during mature biofilm. The first one codes for a putative transferase belonging to the AHBA (3-amino-5-hdroxylbenzoic acid) synthase family, which includes galactosyltransferases involved in the glycosylation of several cell structures like LPS. The second a hypothetical protein-coding gene, but we found that its product is a putative capsule polysaccharide biosynthesis protein that belongs to a family of membrane exporters. Surprisingly, we found that genes involved in the biosynthesis and transport of other common biofilm matrix polysaccharides like alginate were down-regulated in mature biofilm, in particular, the alginate O-acetyltransferase AlgI (*LEPBI_II0277*) and a putative alginate export protein coded by *LEPBI_I3464*. These results indicate that some biofilm components may be produced in the early stages (before 48 h) and their biosynthesis stops once the mature biofilm has been established, while other components seem to be continuously synthesized. Considering this, the biofilm structure is probably being regulated by differential biosynthesis over time, and stages prior to biofilm maturation need to be evaluated in order to decipher which genes are involved in the onset of biofilm formation.

Regarding lipids, we found that key enzymes involved in fatty acid degradation from hexadecanoate were down-regulated in mature biofilms. One of them is a long-chain fatty acid:CoA ligase (EC:6.2.1.3) encoded by *LEPBI_I0107*, that catalyzes the first breaking down step

Table 8.2: Description of differentially expressed genes and biological processes discussed along the manuscript.

Biological process	Gene	Symbol	Description	Biofilm 48 hrs			Biofilm 120 hrs		
				Status	logFC	FDR	Status	logFC	FDR
DNA replication	*LEPBI_I0001*	*dnaA*	Chromosomal replication initiator protein	Down	-0.3	5e-3	-	-	-
	LEPBI_I0012	-	DNA polymerase III, delta subunit	Down	-0.35	3×10^{-3}	-	-	-
	LEPBI_I3461	*dnaX1*	DNA polymerase III, gamma subunit	Down	-0.4	9×10^{-5}	-	-	-
	LEPBI_I3479	*dnaX2*	DNA polymerase III, tau subunit	Down	-0.45	7×10^{-4}	-	-	-
	LEPBI_I3473	*parB*	chromosome partitioning protein ParB	Down	-0.3	4×10^{-3}	-	-	-
	LEPBI_II0026	*parB*	chromosome partitioning protein ParB	Down	-0.39	1×10^{-3}	-	-	-
	LEPBI_I3477	*gidA*	Glucose-imhibited partition protein A	Down	-0.38	5×10^{-4}	-	-	-
	LEPBI_I3475	*gidB*	Glucose-imhibited partition protein B	Down	-0.31	5×10^{-3}	-	-	-
	LEPBI_I0003	*recF*	DNA replication and repair protein RecF	Down	-0.4	1×10^{-4}	-	-	-
	LEPBI_I0005	*gyrB1*	DNA gyrase subunit B	Down	-0.3	2×10^{-3}	-	-	-
Cell growth	*LEPBI_I2249*	*vagC*	Putative virulence-associated protein B	Up	0.9	2×10^{-5}	-	-	-
	LEPBI_I3440	*hepA*	ATP-dependent RNA helicase	Down	-0.4	1×10^{-5}	-	-	-
	LEPBI_I0011	*fis*	Fis family transcriptional regulator	Down	-0.45	4×10^{-3}	Down	-0.5	3×10^{-4}
	LEPBI_I3348	-	Pyrrolo-quinoline quinone	Down	-0.35	3×10^{-4}	-	-	-
Motility	*LEPBI_I0008*	-	PilZ domain	Down	-0.31	9×10^{-3}	-	-	-
	LEPBI_II0088	-	PilZ domain	Down	-0.5	9e-5	Down	-0.44	1×10^{-3}
	LEPBI_I0917	*cheY1*	Chemotactic response regulator CheY	Up	0.47	6×10^{-5}	-	-	-
	LEPBI_I1764	*cheR*	Chemotaxis protein methyltransferase	Down	-0.37	1×10^{-3}	-	-	-
	LEPBI_I1589	*flaB1*	Flagellar filament core protein FlaB	Up	1.19	8×10^{-19}	-	-	-
	LEPBI_I2133	*flaB2*	Flagellar filament 35 kDa core protein	Up	0.66	4×10^{-9}	-	-	-
	LEPBI_I2132	*flaB3*	Flagellar filament 35 kDa core protein	Up	0.98	4×10^{-12}	-	-	-
	LEPBI_I2335	*flaA1*	Flagellar filament outer layer protein A	Up	0.33	6×10^{-3}	-	-	-
Outer membrane proteins	*LEPBI_Ia0817*	*ompL1*	Outer membrane protein OmpL1	Up	0.88	2×10^{-21}	Up	0.31	3×10^{-3}
	LEPBI_I1873	-	OmpA-like protein	Up	0.38	3×10^{-3}	-	-	-
	LEPBI_I0009	-	Putative lipoprotein	Up	0.35	2×10^{-3}	-	-	-
	LEPBI_I1822	-	Putative LipL31	Up	0.38	5×10^{-3}	-	-	-
	LEPBI_2674	*lntB*	Apolipoprotein N-acyltransferase LntB	Up	0.64	8×10^{-9}	-	-	-
Sugar metabolism	*LEPBI_I0073*	*galK*	Galactokinase	Down	-0.48	9×10^{-7}	-	-	-
	LEPBI_I0113	*galE*	Putative UDP-glucose 4-epimerase	Down	-0.32	9×10^{-4}	-	-	-
	LEPBI_I0037	-	Putative transferase	Up	0.37	1×10^{-3}	-	-	-
	LEPBI_I2021	-	Putative capsule polysaccharide biosynthesis protein	Up	0.32	8×10^{-3}	-	-	-
	LEPBI_II0277	*algI*	O-acetyltransferase AlgI	Down	-0.38	5×10^{-4}	-	-	-
	LEPBI_I3464	-	Putative alginate export protein	Down	-0.36	1×10^{-3}	-	-	-
Lipid metabolism	*LEPBI_I0107*	-	Long-chain-fatty-acidŁ??CoA ligase	Down	-0.46	9×10^{-6}	-	-	-
	LEPBI_I0104	*acdA1*	Acyl-CoA dehydrogenase	Down	-0.33	1×10^{-3}	-	-	-
	LEPBI_I0052	-	Enoyl-CoA hydratase	Down	-0.36	4×10^{-4}	-	-	-
	LEPBI_I0777	-	Putative triglyceride lipase	Up	0.69	1.5×10^{-6}	-	-	-
	LEPBI_II0198	*fabG*	3-oxoacyl-ACP reductase	Down	-0.45	7×10^{-7}	-	-	-
	LEPBI_II0199	*fabG*	3-oxoacyl-ACP reductase	Down	-0.39	4×10^{-3}	-	-	-
	LEPBI_II0211	*fabG*	3-oxoacyl-ACP reductase	Down	-0.38	2×10^{-3}	-	-	-
Iron metabolism	*LEPBI_I1883*	*fecA*	Iron(III) dicitrate TonB-dependent receptor	Up	0.5	2×10^{-5}	-	-	-
	LEPBI_I2760	NA	Putative TonB-dependent receptor protein	Up	0.38	1.5×10^{-5}	-	-	-
	LEPBI_I3362	NA	TonB-dependent receptor protein	Down	-0.32	2.7×10^{-3}	-	-	-
	LEPBI_I0669	*hemO*	Heme oxygenase HemO	Up	0.5	1.4×10^{-5}	-	-	-
	LEPBI_p0012	*hemS*	Hemin degradation protein HemS	Up	0.69	5×10^{-4}	-	-	-
	LEPBI_p0015	*hemT*	ABC-type Fe3+-hydroxamate transport system	-	-	-	Up	0.48	3×10^{-4}
	LEPBI_p0014	*hemU*	ABC-type hemin transport system, permease	-	-	-	Up	0.63	8×10^{-7}
	LEPBI_p0013	*hemV*	ABC-type hemin transport system, ATPase	-	-	-	Up	1.1	6×10^{-21}
	LEPBI_I2375	NA	Hemolysin	Up	0,4	1×10^{-3}	-	-	-
Regulators	*LEPBI_I1327*	*flgM*	Anti-sigma factor FlgM	Up	0.7	1×10^{-7}	-	-	-
	LEPBI_I2676	*carD*	CarD family transcriptional regulator	Up	0.86	3×10^{-12}	Up	0.65	1.9×10^{-7}
	LEPBI_I1529	*pnp*	Polynucleotide phsphorylase/polyadenylase	Up	0.38	7×10^{-3}	-	-	-
	LEPBI_I1944	*adk*	Adenylate kinase	Up	1.2	4e-9	Up	0.72	5×10^{-4}
	LEPBI_I1460	*fecR*	FecR protein	Up	0.76	6e-11	Up	0.44	2×10^{-4}
Uncharacterized genes	*LEPBI_I0858*	NA	Putative lipase	Up	1.5	8.5×10^{-9}	-	-	-
	LEPBI_I0859	NA	Putative lipase	Up	1.35	8.5×10^{-30}	-	-	-

of hexadecanoate into acetly-CoA. Additionally, *LEPBI_I0104* (*acdA1*) and *LEPBI_I0052* respectively coding for acyl-CoA dehydrogenase (EC:1.3.8.7, EC:1.3.8.8) and enoyl-CoA hydratase (EC:4.2.1.17) which catalyze downstream enzymatic steps of hexadecanoate degradation, were also down-regulated. Compositional analyses of the biofilm ma-

trix from many bacteria have revealed the presence of fatty acids (hexadecanoic in particular) in abundance [300–303]. Hence, the accumulation of hexadecanoate by down-regulating the enzymes involved in fatty acid degradation may indicate that these lipids could be used for matrix composition in *L. biflexa*. This hypothesis is reinforced by the over-expression of *LEPBI_I0777* which codes for a triglyceride lipase (EC:3.1.1.3), allocated to glycerolipid metabolism and involved in degrading triglycerides to single fatty acids (Tab. 8.2). Furthermore, most enzymes belonging to the canonical pathway of fatty acid biosynthesis from acetyl-CoA were not differentially expressed during biofilm formation, suggesting that fatty acids are being synthesized *de novo* during biofilm. Exceptionally, we found one *fabG* homolog (*LEPBI_II0198*) down-regulated. FabG (EC:1.1.1.100) catalyzes a key step of fatty acid biosynthesis so its depletion should indicate reduced or null production of fatty acids. Further analysis of *L. biflexa* genome revealed the presence of 16 putative *fabG* homologs with total conservation of catalytic residues but just 3 out of them (*LEPBI_II0198*, *LEPBI_II0199* and *LEPBI_II0211*) were down-regulated. Beyond their role in biofilm formation, the great level of paralogy for this gene, unique among the genes in the pathway, may suggest functional redundancy or the evolution of substrate-specific FabG isoforms. The fact that only 3 *fabG* copies were down-regulated may indicate that the fatty acid biosynthetic pathway is not stopped at FabG step, but further investigation will be needed for elucidating the striking role of the high redundancy of *fabG* in *L. biflexa* biology.

8.4.7 Iron uptake. Iron plays a crucial role in biological processes by composing essential enzyme cofactors or in electron transport chains. *Leptospira* require an environmental iron source to grow and, as many other bacteria, have evolved diverse strategies to scavenge it from its surroundings. Considering that biofilm-embedded cells are mostly sessile, we thought that these scavenging systems should be transcriptionally altered in response to iron availability. The genomic and functional characterization of iron uptake systems in *L. biflexa* have revealed the presence of redundant outer membrane TonB-dependent receptors (with different specificities for iron containing compounds), hemolysins, inner membrane hemin transporters and the FeoAB sys-

tem, but an absence of genes coding for siderophore biosynthesis [304].

A salient feature of mature biofilm was the over-expression of *LEPBI_I1883*, coding for *fecA*. Functional characterization of FecA in *L. biflexa* revealed its capacity to transport diverse iron compounds like aerobactin, iron citrate, iron chloride and iron sulfate [304]. Interestingly, the EMJH medium where *L. biflexa* were cultured in this work contains iron sulfate as an iron source, suggesting that the sessile condition of biofilm cells requires the up-regulation of this TonB-dependent receptor to encompass iron acquisition. Another TonB-dependent receptor-coding gene (*LEPBI_I2760*) was also over-expressed in mature biofilm, whose disruption impairs the ability to use desferrioxamine as iron source in *L. biflexa* [304]; and a similar TonB-dependent receptor (FoxA) is responsible for desferrioxamine utilization in *Yersinia enterocolitica* [305]. Furthermore, an additional TonB-dependent receptor-coding gene (*LEPBI_I3362*) was down-regulated in mature biofilm. Disrupting *LEPBI_I3362* leads to a wild-type phenotype in iron-depleted medium probably due to functional redundancy with other iron uptake systems [304]. Translocation of iron compounds from the periplasmic space to the cytoplasm in *L. biflexa* relies on siderophore- or metal-ABC transporters, the FeoAB system and the hemin uptake system. Except for the hemin uptake system, none of these transporters were differentially expressed during biofilm formation. We also found that both genes coding for the heme oxygenase HemO (*LEPBI_I0669*) and the hemin degradation protein HemS (*LEPBI_p0012*) were up-regulated; however, differential expression of the remaining ABC transporter components HemT, HemU and HemV was only detected in late biofilm. The gene *LEPBI_I2375* that codes for a hemolysin was also up-regulated. These results evidence that iron uptake is finely tuned during biofilm formation, considering that only some specific TonB-dependent receptors, the hemin uptake system and one hemolysin were differentially expressed in this experiment.

8.4.8 Regulatory genes and co-regulation networks. In the previous sections we have presented and discussed the most relevant protein-coding genes and gene pathways that we found altered when comparing biofilms with planktonic cells in *L. biflexa*, omitting how

these genes can be modulated through the action of other regulatory genes. To assess this, we recovered all differentially expressed genes involved in any regulatory step, like transcription factors, and investigated how their transcription levels co-varied with the rest of the differentially expressed genes in order to describe co-regulation patterns among them. Out of 575 differentially expressed genes in any condition, 47 (8%) were annotated as transcription factors or related proteins involved in regulatory processes. In general, regulators can be classified as activators or repressors if they enhance or reduce the transcription of other genes. In this sense, we found that most regulatory networks were conformed by genes whose transcription levels correlated positively, which suggests most relevant regulatory processes were orchestrated by activators (Fig. 8.2). Out of the 13 different co-activation networks that were identified, 5 (accounting for 64% of co-regulated genes) were differentially expressed in mature biofilm while 8 (accounting for 36% of co-regulated genes) were differentially expressed in late biofilm, evidencing that around 2/3 of co-regulatory processes are taking place in mature biofilm.

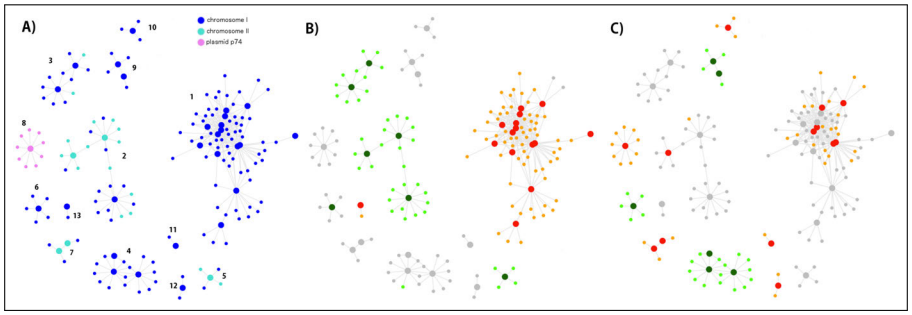


Figure 8.2: Co-expression networks. This figure shows the 13 co-expression networks that resulted from analyzing positively correlated genes. The big circles represent genes involved in regulatory processes and are colored in red if they are up-regulated and in dark green if they are down regulated. Small circles are colored in orange for up-regulated and light green for down-regulated genes. Grey circles are genes without differential expression in that condition. A) Genes are colored by replicon. B) Differentially expressed genes at 48 hrs. C) Differentially expressed genes at 120 hrs.

We identified a predominant co-regulation network that alone includes 40% of co-regulated genes and also contains regulators that direct most relevant functions for biofilm; a detailed description of this

network is presented in (Fig. 8.3). The regulator with highest number of interactions (co-expressed genes) was *LEPBI_I1327*, a hypothetical protein coding gene. However, more careful analyses (see Methods) revealed it codes for the anti-sigma factor FlgM, which interacts with sigma factor FliA. The presence of FliA and flagellar components FlaA1 and FlaB123 in the network suggested this system may be responsible for tuning flagellar function, hence bacterial motility. Additionally, the presence of 5 different anti-anti-sigma factors in the network implies that such antagonists also regulate anti-sigma factors like FlgM. This reveals a previously unreported role of the anti-anti-sigma regulatory system in the physiology of *L. biflexa*. However, further experimental work will be needed to confirm our structural annotation and to demonstrate that *LEPBI_I1327* codes for FlgM and it regulates FliA.

The second regulator in the ranked interactions was *LEPBI_I2676*, encoding a transcription factor resembling mycobacterial CarD, which is over-expressed during hostile conditions like nutrient deprivation [306]. Interestingly, the gene coding for the regulatory protein polynucleotide phosphorylase (Pnp, *LEPBI_I1529*) was co-expressed with *LEPBI_I2676*, and a functional relationship between these two genes has been shown in *Mycobacterium tuberculosis* [306]. Furthermore, a specific role for Pnp during biofilm formation has been established in *Salmonella*, where the expression of CsgD (the master activator of biofilm formation) was substantially reduced in the pnp mutant [307]. The same study also set Pnp as an indirect regulator of cyclic monophosphate nucleotides, which are key second messengers in biofilm formation. In this sense, we found that *adk* gene (*LEPBI_I1944*), coding for an adenlyate/guanylate cyclase, was co-expressed with *pnp*. The underlying molecular mechanism for the regulatory role of Pnp may involve its RNAse activity that selectively degrades mRNAs [308]. This incorporates a post-transcriptional regulation step and would put this protein as a key modulator of genes involved in *L. biflexa* biofilms.

Another top-scoring co-regulator was FecR (*LEPBI_I1460*), which is needed for the induction of *fecABCDE* iron transport operon in enterobacteria like *E. coli* [309]. Interestingly, FecA (*LEPBI_I1883*) is present in this co-regulation network, suggesting an important role for this gene tandem in iron acquisition. Furthermore, previous studies were unable to identify *fecBCDE* homologs in *L. biflexa* suggesting a differ-

ent unknown pathway for these functions [304]. The study of the 11 co-expressed genes with FecR, now annotated as hypothetical proteins with remote or no homology in sequence databases, will probably shed light on unknown aspects of iron metabolism in *L. biflexa*.

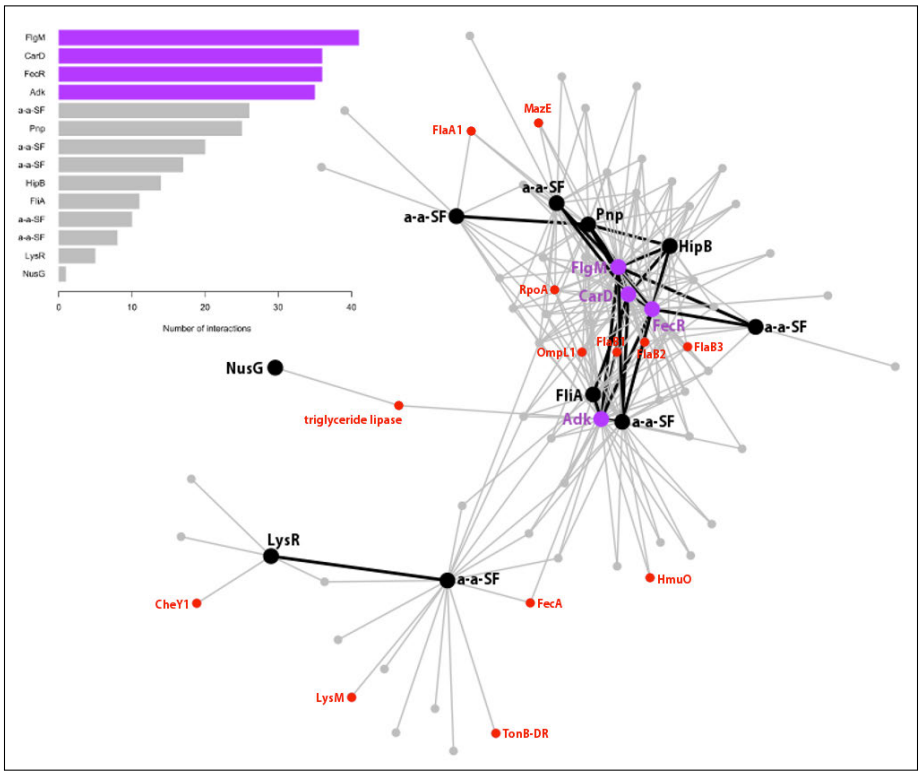


Figure 8.3: Major co-expression network. This figure shows the biggest co-expression network (1). The barplot shows all regulatory genes in the network sorted by the number of interactions (co-expressed genes). The top 4 regulatory genes are highlighted in purple, while the rest are highlighted in black. Other relevant genes that were discussed along the manuscript are in red. Black edges represent co-expression between regulatory genes.

When analyzing where in CI, CII or P74 the regulators and their cognate genes were coded, we found that for a given regulatory network, almost all genes linked to it were coded in the same replicon. However, Fig. 8.2 also shows that a minority of genes from CI and CII are co-expressed in the same network. This kind of inter-chromosomal regulation has been evidenced in other bacteria with multiple replicons

like *Vibrio cholerae* [310]. At any rate, these findings support the hypothesis introduced previously, namely that each replicon in *L. biflexa* plays particular roles during biofilm formation, with minimal interaction between gene products from distinct replicons.

8.4.9 Small regulatory RNAs. Small regulatory RNAs. Small RNA regulators (sRNAs) have been identified in a wide range of organisms including bacteria, and found to play important regulatory roles in several biological processes [311]. Recently, a paper describing the transcriptional adaptation of *L. interrogans* to the intra-host environment has evidenced the expression of sRNAs in this species [312]. The transcription of non-coding regions with sRNA signatures has not been reported so far in *L. biflexa*. Using de novo prediction tools we identified 181 putative sRNAs dispersed in the 3 replicons (CI = 168, CII = 11, P74 = 2) and just 30 out of them (15%) showed to be transcriptionally active (CPM > 1); these active sRNAs were placed in CI (n = 25) and CII (n = 5) (Supp. Tab. 8.5).

Among others, one anti-sense sRNA of 93-bp placed in CII next to the alginate biosynthesis genes was down-regulated in mature biofilm (logFC = -0.45, FDR = 4×10^{-4}). Strikingly, the same sRNA was up-regulated in late biofilm (logFC = 0.58, FDR = 1.4×10^{-4}). It worth mentioning that the alginate O-acetyltransferase coding gene *algI* was down-regulated in mature biofilm and unchanged in late biofilm. Further characterization of this and others candidate sRNAs is required to understand their role in the regulation of genes involved in biofilm formation.

8.4.10 Differentially expressed genes of unknown function. The phylum Spirochaetes has evolved many distinctive and often intriguing features since its deep branching in the bacterial phylogeny. Accordingly, a great number of leptospiral genes code for hypothetical proteins with limited or null homology in sequence databases, challenging downstream experimental procedures based on predicted protein functions. In this RNA-seq experiment, we found that 289 out of 575 (50%) differentially expressed genes in any condition were annotated as hypothetical protein-coding genes. Even after refined manual curation, more than 50 differentially expressed genes remained with-

out any predicted function. Indeed, among the top 5 up-regulated genes (ranked by fold change) in mature biofilm we found two consecutive genes (*LEPBI_I0858*, *LEPBI_I0859*) that were originally annotated as hypothetical proteins; however, structural annotation revealed that they probably have a lipase activity. This is a strong evidence that genes encoding hypothetical proteins in *Leptospira* are true and actively transcribed genes whose functions remain to be determined, opening new grounds of research in leptospiral biology. Structural annotations are presented in Supp. Tab. 8.6.

8.4.11 RT-PCR confirmation of selected genes. In order to check the robustness and reproducibility of differentially expressed genes detected by the RNA-seq analysis, a set of 21 genes was used to perform relative quantification by RT-PCR. These genes are representative for the most relevant pathways discussed along the manuscript. Supp. Fig. **??** shows that for the vast majority of tested genes, the expression levels were coherent with those observed throughout RNA-seq analysis and differences were statistically significant ($P < 0.05$, T-test). Furthermore, RT-PCRs were performed with a set of template RNAs derived from an independent biofilm experiment than the one used for performing RNA-seq, indicating significant reproducibility of detected transcript switches in these genes. Additionally, we proposed a set of *L. biflexa* genes that can be used for RT-PCR normalization due to their scarce transcription variability along biofilm and planktonic states (Supp. Tab. 8.3).

8.4.12 Integrative view of gene expression in biofilm formation. In this work we describe the first RNA-seq experiment performed over the model organism *L. biflexa* oriented to gene expression changes in biofilms over abiotic surfaces at two time points (48 h and 120 h). At 48 h *L. biflexa* reaches an optimal biofilm growth denoted as mature [267], that when compared to planktonic state at the same time allowed to identify chief genetic factors differentiating biofilm from free-living states. At 120 h the (late) biofilm structure presents signs of cell disaggregation, evidencing the known detachment process responsible for biofilm-to-planktonic cells recycling or even cell death. In this sense, some genetic changes associated to this process could be

identified, such as expression levels of several genes involved in adhesion and EPS production. However, and despite our paper describes a clear transcription turnover between biofilm and planktonic states, performing transcriptomics on biofilm cultures previous to 48 h will contribute to identify additional features that determine biofilm formation and, in particular, those genes that govern the initial phase of interaction between cells and the abiotic surface.

As an outline, our results highlighted many functions related to cell growth and metabolism that were altered during biofilm, DNA replication and cell division probably being the most notable. Additionally, other well-recognized pathways like sugar, lipid and iron metabolism presented transcriptional regulation. Beyond identifying the role of these well-known metabolic pathways during biofilm formation, we made annotation improvements for many genes lacking assigned molecular function. In this sense, a most remarkable case is *LEBI_I1327*, which was previously annotated as a hypothetical protein but we propose it as homologous to *flgM*. Indeed, it seems to be the most relevant regulatory gene during biofilm formation based on our co-expression analysis. Furthermore, we reported for first time in *L. biflexa* the presence of sRNA regulators that were transcriptionally active during biofilm and planktonic growth. Despite identifying some candidate sRNA for the regulation of particular processes like alginate biosynthesis, a more detailed and specific work centered in the analysis of sRNAs is required to understand their targets and regulation networks.

Another hint shown by our results is a possible differentiation between biofilm formation mechanisms in saprophytic and pathogenic leptospires, as testified by the presence of differentially expressed genes probably involved in cell-to-cell adhesion that were not found in the genomes of pathogenic *Leptospira*. In particular, as the pathogenic *L. interrogans* shares with *L. biflexa* the free-living trait, where probably biofilm formation is crucial for bacterial persistence, elucidating transcriptional changes in *L. interrogans* during biofilm formation would be very informative for determining if both species have evolved particular features associated to this growing condition. Finally, taking into account the recent availability of genomic sequences for several leptospiral species with differential incidence in leptospirosis, we con-

sider that further extending RNA-seq analyses to species belonging to Group I, II and III could shed more light on the evolution of this striking organisms, as well as contributing to generate effective tools for leptospirosis control.

8.5 Acknowledgements

We thank Mitermayer Galvão dos Reis and Clarissa Araújo Silva Gurgel for kindly providing strains and reagents. This work was supported by the Agencia Nacional de Investigación e Innovación (ANII, Uruguay), grant FCE_2_2011_1_7179; the Coordenação de Aperfeçioamento de Pessoal de Nivel Superior (CAPES, Brazil), grant 034/2012; and by the Mercosur Structural Convergence Fund (FOCEM), grant COF 03/11. GI also thanks to the Comisión Sectorial de Investigación Científica (CSIC, Uruguay) for support.

8.6 References

[258] B. Abela-Ridder, R. Sikkema, R. A. Hartskeerl, *Int. J. Antimicrob. Agents* **2010**, *36 Suppl 1*, 5–7.

[259] M. Picardeau, D. M. Bulach, C. Bouchier, R. L. Zuerner, N. Zidane, P. J. Wilson, S. Creno, E. S. Kuczek, S. Bommezzadri, J. C. Davis, A. McGrath, M. J. Johnson, C. Boursaux-Eude, T. Seemann, Z. Rouy, R. L. Coppel, J. I. Rood, A. Lajus, J. K. Davies, C. Medigue, B. Adler, *PLoS ONE* **2008**, *3*, e1607.

[260] B. Adler, *Curr. Top. Microbiol. Immunol.* **2015**, *387*, 1–9.

[261] J. S. Lehmann, M. A. Matthias, J. M. Vinetz, D. E. Fouts, *Pathogens* **2014**, *3*, 280–308.

[262] L. Hall-Stoodley, P. Stoodley, *Trends Microbiol.* **2005**, *13*, 7–10.

[263] M. Guell, E. Yus, M. Lluch-Senar, L. Serrano, *Nat. Rev. Microbiol.* **2011**, *9*, 658–669.

[264] A. Dotsch, D. Eckweiler, M. Schniederjans, A. Zimmermann, V. Jensen, M. Scharfe, R. Geffers, S. Haussler, *PLoS ONE* **2012**, *7*, e31092.

[265] S. A. Frese, D. A. Mackenzie, D. A. Peterson, R. Schmaltz, T. Fangman, Y. Zhou, C. Zhang, A. K. Benson, L. A. Cody, F. Mulholland, N. Juge, J. Walter, *PLoS Genet.* **2013**, *9*, e1004057.

[266] S. Rumbo-Feal, M. J. Gomez, C. Gayoso, L. Alvarez-Fraga, M. P. Cabral, A. M. Aransay, N. Rodriguez-Ezpeleta, A. Fullaondo, J. Valle, M. Tomas, G. Bou, M. Poza, *PLoS ONE* **2013**, *8*, e72968.

[267] P. Ristow, P. Bourhy, S. Kerneis, C. Schmitt, M. C. Prevost, W. Lilenbaum, M. Picardeau, *Microbiology (Reading Engl.)* **2008**, *154*, 1309–1317.

[268] B. Brihuega, L. Samartino, C. Auteri, A. Venzano, K. Caimi, *Rev. Argent. Microbiol.* **2012**, *44*, 138–143.

[269] Y. Liao, G. K. Smyth, W. Shi, *Nucleic Acids Res.* **2013**, *41*, e108.

[270] M. D. Robinson, D. J. McCarthy, G. K. Smyth, *Bioinformatics* **2010**, *26*, 139–140.

[271] M. J. Cros, A. de Monte, J. Mariette, P. Bardou, B. Grenier-Boley, D. Gautheret, H. Touzet, C. Gaspin, *RNA* **2011**, *17*, 1947–1956.

[272] J. Soding, *Bioinformatics* **2005**, *21*, 951–960.

[273] M. Remmert, A. Biegert, A. Hauser, J. Soding, *Nat. Methods* **2012**, *9*, 173–175.

[274] G. Csardi, T. Nepusz, *InterJournal Complex Systems* **2006**, *1695*, 1–9.

[275] Z. Wang, M. Gerstein, M. Snyder, *Nat. Rev. Genet.* **2009**, *10*, 57–63.

[276] B. J. Haas, M. Chin, C. Nusbaum, B. W. Birren, J. Livny, *BMC Genomics* **2012**, *13*, 734.

[277] G. Nucifora, L. Chu, S. Silver, T. K. Misra, *J. Bacteriol.* **1989**, *171*, 4241–4247.

[278] J. U. Kreft, *Microbiology (Reading Engl.)* **2004**, *150*, 2751–2760.

[279] E. Aizenman, H. Engelberg-Kulka, G. Glaser, *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 6059–6063.

[280] K. Pedersen, S. K. Christensen, K. Gerdes, *Mol. Microbiol.* **2002**, *45*, 501–510.

[281] S. Karlin, J. Mrázek, A. Campbell, D. Kaiser, *Journal of bacteriology* **2001**, *183*, 5025–5040.

[282] M. D. Bradley, M. B. Beach, A. J. de Koning, T. S. Pratt, R. Osuna, *Microbiology* **2007**, *153*, 2922–2940.

[283] N. W. Charon, G. R. Daughtry, R. S. McCuskey, G. N. Franz, *J. Bacteriol.* **1984**, *160*, 1067–1073.

[284] X. Fang, M. Gomelsky, *Molecular microbiology* **2010**, *76*, 1295–1305.

[285] Z.-H. Li, K. Dong, J.-P. Yuan, B.-Y. Hu, J.-X. Liu, G.-P. Zhao, X.-K. Guo, *Biochemical and biophysical research communications* **2006**, *345*, 858–866.

[286] M. Picardeau, A. Brenot, I. Saint Girons, *Molecular microbiology* **2001**, *40*, 189–199.

[287] A. Lambert, M. Picardeau, D. A. Haake, R. W. Sermswan, A. Srikram, B. Adler, G. A. Murray, *Infection and immunity* **2012**, *80*, 2019–2025.

[288] D. A. Haake, M. K. Mazel, A. M. McCoy, F. Milward, G. Chao, J. Matsunaga, E. A. Wagar, *Infection and immunity* **1999**, *67*, 6572–6582.

[289] L. G. Fernandes, M. L. Vieira, K. Kirchgatter, I. J. Alves, Z. M. de Morais, S. A. Vasconcellos, E. C. Romero, A. L. Nascimento, *Infection and immunity* **2012**, *80*, 3679–3692.

[290] J. K. Barnett, D. Barnett, C. A. Bolin, T. A. Summers, E. A. Wagar, N. F. Cheville, R. A. Hartskeerl, D. A. Haake, *Infect. Immun.* **1999**, *67*, 853–861.

[291] R. Orme, C. W. Douglas, S. Rimmer, M. Webb, *Proteomics* **2006**, *6*, 4269–4277.

[292] P. A. Cullen, D. A. Haake, D. M. Bulach, R. L. Zuerner, B. Adler, *Infection and immunity* **2003**, *71*, 2414–2421.

[293] J. E. Nally, J. P. Whitelegge, S. Bassilian, D. R. Blanco, M. A. Lovett, *Infection and immunity* **2007**, *75*, 766–773.

[294] C. Robichon, D. Vidal-Ingigliardi, A. P. Pugsley, *Journal of Biological Chemistry* **2005**, *280*, 974–983.

[295] M. Moreno-Paz, M. J. Gomez, A. Arcas, V. Parro, *BMC Genomics* **2010**, *11*, 404.

[296] Y. Chai, P. B. Beauregard, H. Vlamakis, R. Losick, R. Kolter, *MBio* **2012**, *3*, e00184–12.

[297] T. Taniyama, Y. Yanagihara, I. Mifuchi, I. Azuma, Y. Yamamura, *Infection and immunity* **1972**, *6*, 414.

[298] D. M. Bulach, T. Kalambaheti, A. de la Peña-Moctezuma, B. Adler, *Infection and immunity* **2000**, *68*, 3793–3798.

[299] C. Whitfield, *Trends in microbiology* **1995**, *3*, 178–185.

[300] M. Timke, D. Wolking, N. Q. Wang-Lieu, K. Altendorf, A. Lipski, *Applied microbiology and biotechnology* **2004**, *66*, 100–107.

[301] C. Theilacker, P. Sanchez-Carballo, I. Toma, F. Fabretti, I. Sava, A. Kropec, O. Holst, J. Huebner, *Molecular microbiology* **2009**, *71*, 1055–1069.

[302] L Bruno, F Di Pippo, S Antonaroli, A Gismondi, C Valentini, P Albertano, *Journal of applied microbiology* **2012**, *113*, 1052–1064.

[303] L. Purish, L. Asaulenko, D. Abdulina, V. VasilâĂŹev, G. Iutinskaya, *Applied biochemistry and microbiology* **2012**, *48*, 262–269.

[304] H Louvel, S Bommezzadri, N Zidane, C Boursaux-Eude, S Creno, A Magnier, Z Rouy, C Medigue, I Saint Girons, C Bouchier, et al., *Journal of bacteriology* **2006**, *188*, 7893–7904.

[305] A. J. Báumler, K. Hantke, *Molecular microbiology* **1992**, *6*, 1309–1321.

[306] C. L. Stallings, N. C. Stephanou, L. Chu, A. Hochschild, B. E. Nickels, M. S. Glickman, *Cell* **2009**, *138*, 146–159.

[307] S. F. Rouf, I. Ahmad, N. Anwar, S. K. Vodnala, A. Kader, U. Römling, M. Rhen, *Journal of bacteriology* **2011**, *193*, 580–582.

[308] K. Yamanaka, M. Inouye, *Journal of Bacteriology* **2001**, *183*, 2808–2816.

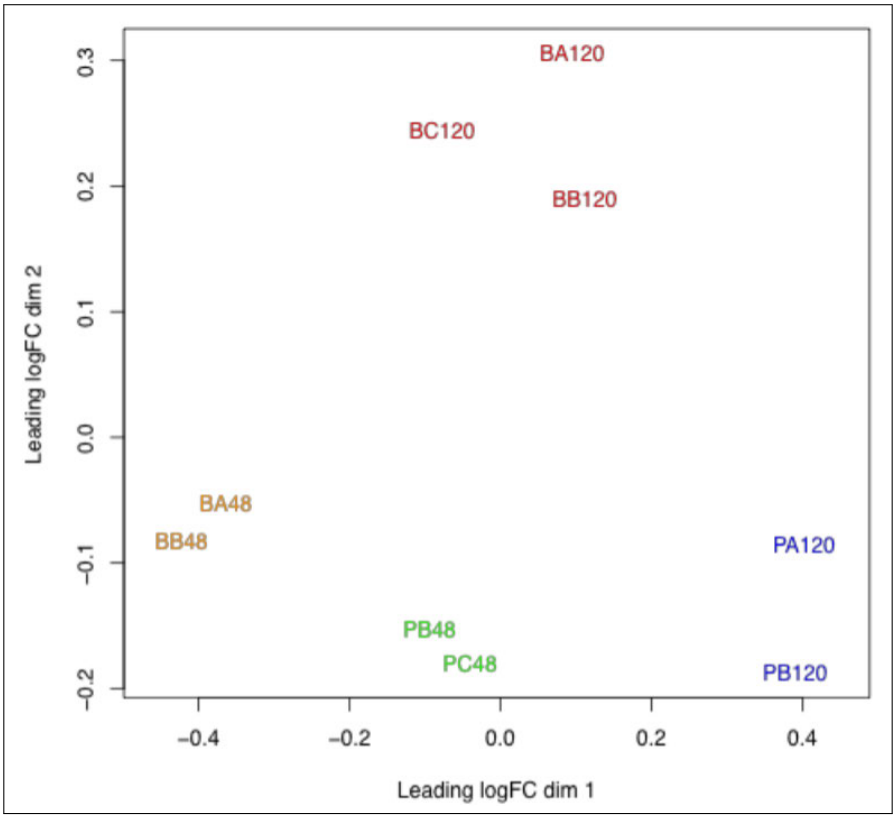
[309] M. Ochs, S. Veitinger, I. Kim, D. Weiz, A. Angerer, V. Braun, *Molecular microbiology* **1995**, *15*, 119–132.

[310] F. H. Yildiz, G. K. Schoolnik, *Journal of bacteriology* **1998**, *180*, 773–784.

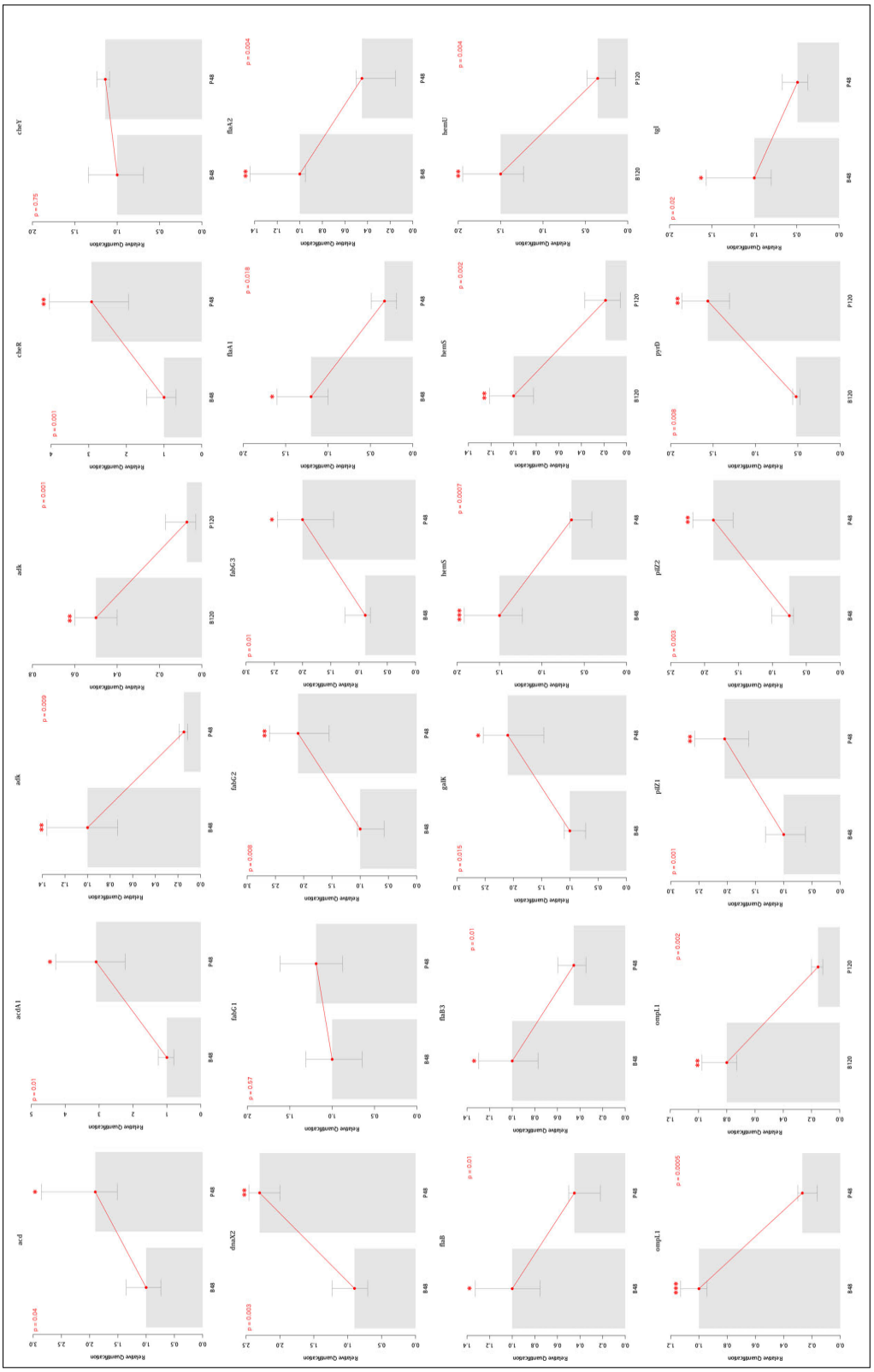
[311] S. Gottesman, G. Storz, *Cold Spring Harbor perspectives in biology* **2011**, *3*, a003798.

[312] M. J. Caimano, S. K. Sivasankaran, A. Allard, D. Hurley, K. Hokamp, A. A. Grassmann, J. C. Hinton, J. E. Nally, *PLoS Pathog* **2014**, *10*, e1004004.

8.7 Supplementary material



Supp. Fig. 8.1: MDS plot. Samples considered in differential expression analysis.



Supp. Fig. 8.2: RT-PCR analysis. Relative quantification for selected genes. Asterisks show statistical significance assessed by t-test ($p < 0.05$).

Supp. Tab. 8.1: Information for deposited data at Sequence Read Archive.

BioProject	Accession	Sample Name	Organism	Strain	TaxID
PRJNA288909	SAMN04364752	BA48	Leptospira biflexa serovar Patoc	Patoc I (Paris)	172
PRJNA288909	SAMN04364753	BB48	Leptospira biflexa serovar Patoc	Patoc I (Paris)	172
PRJNA288909	SAMN04364754	BC48	Leptospira biflexa serovar Patoc	Patoc I (Paris)	172
PRJNA288909	SAMN04364755	PA48	Leptospira biflexa serovar Patoc	Patoc I (Paris)	172
PRJNA288909	SAMN04364756	PB48	Leptospira biflexa serovar Patoc	Patoc I (Paris)	172
PRJNA288909	SAMN04364757	PC48	Leptospira biflexa serovar Patoc	Patoc I (Paris)	172
PRJNA288909	SAMN04364758	BA120	Leptospira biflexa serovar Patoc	Patoc I (Paris)	172
PRJNA288909	SAMN04364759	BB120	Leptospira biflexa serovar Patoc	Patoc I (Paris)	172
PRJNA288909	SAMN04364760	BC120	Leptospira biflexa serovar Patoc	Patoc I (Paris)	172
PRJNA288909	SAMN04364761	PA120	Leptospira biflexa serovar Patoc	Patoc I (Paris)	172
PRJNA288909	SAMN04364762	PB120	Leptospira biflexa serovar Patoc	Patoc I (Paris)	172
PRJNA288909	SAMN04364763	PC120	Leptospira biflexa serovar Patoc	Patoc I (Paris)	172

Supp. Tab. 8.2: Primers used in RT-PCR analysis.

Gene	Name	Sequence (5' ->3')	Strand	Bp	Tm	GC
LEPBI_I0073	galK-F	AGTGGTAACTGGCTTTGCGA	Plus	20	59.89	50
	galK-R	GCTTTCTGTCCAATCACGGC	Minus	20	59.83	55
LEPBI_I3479	dnaX2-F	AGAACCGCCTCCACATACAA	Plus	20	59.02	50
	dnaX2-R	CTGCGGGAGGTAGTGGAAAG	Minus	20	60.11	60
LEPBI_I1589	flaB-F	GTCTAACGACGCGAACCTGA	Plus	20	60.11	55
	flaB-R	CTGCAAGTCCAGATGCGTCA	Minus	20	60.67	55
LEPBI_I1944	adk-F	CAAGCAAAGGCTCTCTCGGA	Plus	20	59.75	50
	adk-R	AGCGTCCTTCTTTGATCGCT	Minus	20	59.61	55
LEPBI_Ia0817	ompL1-F	AGTGGGTTCGGTCTCAACTG	Plus	20	59.61	55
	ompL1-R	GAGCAGAAGCTCCACCGATT	Minus	20	60.11	55
LEPBI_I2132	flaB3-F	GCAAACGCAAGGCAAAGAGA	Plus	20	59.97	50
	flaB3-R	TTTCCAGGTGTCGAGAGTGC	Minus	20	59.97	55
LEPBI_p0012	hemS-F	AATTCGAGACGCAGCCAAAC	Plus	20	59.48	50
	hemS-R	CCCAAGTTTTGGCGTTTCCA	Minus	20	59.54	50
LEPBI_p0014	hemU-F	AGTTTGGGAGGGGCATCTTG	Plus	20	59.96	55
	hemU-R	CCAAGTGACCTGCTTCTCGT	Minus	20	59.97	55
LEPBI_I0092	pyrD-F	GACTTGCCGCTGGATTTGAC	Plus	20	59.83	50
	pyrD-R	TTTGGCTGTGATGGTTCCGA	Minus	20	59.89	55
LEPBI_I0008	pilZ1-F	GACTAGCCTTTCAAACGACA	Plus	20	55.72	45
	pilZ1-R	AGAGGTTTGAAAATCACCGA	Minus	20	54.89	40
LEPBI_II0088	pilZ2-F	GTATCCAAAGGCAAAAGTGG	Plus	20	54.84	45
	pilZ2-R	TATCGTCCTCAAAAAGTTGGT	Minus	21	55.16	38
LEPBI_I0917	cheY1-F	GGTATGACGGGAATCGAATTA	Plus	21	55.29	43
	cheY1-R	GGTTTTACAAGCCAACCAAC	Minus	20	55.66	45
LEPBI_I1764	cheR-F	TTACTCCAGTTTCCGTTTCC	Plus	20	55.31	45
	cheR-R	AGGATCAAATACCCTTTGGG	Minus	20	54.65	45
LEPBI_I2335	flaA1-F	TGAATCTTGGGACAATCCAG	Plus	20	55.02	45
	flaA1-R	GATTTTGCTGGGTCATTCAG	Minus	20	54.93	45
LEPBI_I2336	flaA2-F	ACAGACACACCTTATTTGCT	Plus	20	54.88	40
	flaA2-R	TTGCTGTCAACTTTCTCCAT	Minus	20	55.17	40
LEPBI_I0104	acdA1-F	AAGAGTATGGTGGTATGGGT	Plus	20	55.19	45
	acdA1-R	TCCTTGTTGTGCTTGGATTA	Minus	20	54.86	40
LEPBI_I0052	acd-F	ATGAGAGACCTTGGTGAGAT	Plus	20	55.23	45
	acd-R	TTTCTGCATCCAATCCGTTA	Minus	20	55.05	40
LEPBI_I0777	tgl-F	TTTTTAGCGACCCTTCTCTC	Plus	20	55.1	45
	tgl-R	CCTCCCCAATACTTTACGAG	Minus	20	54.95	50
LEPBI_II0198	fabG1-F	AAGGATTCGATTGTTCTCGT	Plus	20	54.8	40
	fabG1-R	GGTTTCCTGTAGAATGGGTT	Minus	20	54.92	45
LEPBI_II0199	fabG2-F	CGAACTATCTCTTGCTGGAA	Plus	20	54.31	45
	fabG2-R	TACACAATGAGTTCTGGACG	Minus	20	55.14	45
LEPBI_II0211	fabG3-F	CGAAGAACTTGCCATTTACC	Plus	20	54.99	45
	fabG3-R	ACACGATGGAGGATACAATC	Minus	20	54.65	45
LEPBI_I2771	LEPBI_I2771-F	CTCTCGGTGGAGTTTTCGGT	Plus	20	59.68	55
	LEPBI_I2771-R	AACAAATCCCTTCGCCAGCA	Minus	20	60.64	50

Supp. Tab. 8.3: Genes used for normalization in RT-PCR relative quantification.

Gene	Product	CPM								
		BA120	BA48	BB120	BB48	BC120	PA120	PB120	PB48	PC48
LEPBI_I1415	hypothetical protein	439.3144	458.2669	446.4948	433.2076	453.9055	467.2638	433.2911	442.8527	424.992
LEPBI_I1808	hypothetical protein	472.9315	476.9131	473.6058	486.5844	462.4646	492.072	459.8609	465.6223	480.1459
LEPBI_I2349	Mrp family ATP-binding protein	372.4622	386.1203	381.2275	373.9328	380.1873	359.5911	389.1445	371.2439	379.0303
LEPBI_I2735	phpshoribosylglycinamide formyltransferase 2	332.7329	344.6683	338.0508	327.0437	360.0321	339.898	360.1222	351.7742	334.6008
LEPBI_I2771	nitrite extrusion protein 1 NarK	503.8746	479.0646	477.9569	478.6221	495.3203	485.9339	490.5183	474.5322	471.8729
LEPBI_I3250	hypothetical protein	343.0473	357.7207	343.0713	327.0437	354.7863	342.4555	350.3118	350.7842	337.3584

Supp. Tab. 8.4: Reads mapped by sample.

Sample	Total reads	Mapped reads	Percentage mapped
BA120	2804203	2782554	99.23
BA48	7500998	7411156	98.8
BB120	3207624	3174615	98.97
BB48	3642931	3603771	98.93
BC120	3875081	3847960	99.3
BC48	3803957	3770114	99.11
PA120	4203432	4150680	98.75
PA48	2988218	2947702	98.64
PB120	2631490	2598663	98.75
PB48	3248509	3217719	99.05
PC120	3963865	3921581	98.93
PC48	3495242	3465387	99.15

Supp. Tab. 8.5: Expression of predicted sRNAs.

Chromosome	Start	End	Length	Strand	Description	Active	Differentially expressed (logFC)			
							BvsP_48	BvsP_120	B_120vs48	P_120vs48
chrI	2043554	2043715	162	-	Lysine	No	-	-	-	-
chrI	2806361	2806534	174	+	Lysine	No	-	-	-	-
chrI	246415	246456	42	-	SAM	No	-	-	-	-
chrI	648107	648163	57	-	RtT	No	-	-	-	-
chrI	2380114	2380168	55	-	RtT	No	-	-	-	-
chrI	2468926	2469054	129	-	LR-PK1	No	-	-	-	-
chrI	414383	414453	71	+	CAESAR	Yes	-	-	-	-
chrI	796163	796191	29	+	RprA	Yes	-	-	-	-
chrI	877669	877699	31	-	RprA	Yes	-	-0,70	-	1,3
chrI	3473660	3473689	30	-	RprA	Yes	-	-	-	-
chrI	906252	906279	28	-	SraE/RygA/RygB_family	No	-	-	-	-
chrI	1628886	1628914	29	-	SraE/RygA/RygB_family	No	-	-	-	-
chrI	2559248	2559275	28	-	RyeE	Yes	-	-	-	-
chrI	343749	343817	69	-	SL2	No	-	-	-	-
chrI	1155497	1155531	35	-	Threonine_leader	Yes	-	-	-	-
chrI	906256	906281	26	+	Leucine_leader	No	-	-	-	-
chrI	2407469	2407494	26	+	Leucine_leader	Yes	-	-	-	-
chrI	906256	906281	26	-	Leucine_leader	No	-	-	-	-
chrI	2407469	2407494	26	-	Leucine_leader	No	-	-	-	-
chrI	2778262	2778283	22	+	Pseudomonas_sRNA_P9	No	-	-	-	-
chrI	3297300	3297322	23	+	Pseudomonas_sRNA_P9	No	-	-	-	-
chrI	429033	429133	101	+	GEMM_cis-regulatory_element	No	-	-	-	-
chrI	2620243	2620310	68	-	Pseudoknot	No	-	-	-	-
chrI	2922549	2922616	68	-	Pseudoknot	No	-	-	-	-
chrI	2429451	2429479	29	+	CRISPR_repeat	No	-	-	-	-
chrI	3427282	3427302	21	+	CRISPR_repeat	No	-	-	-	-
chrI	355359	355369	11	+	CRISPR_repeat	No	-	-	-	-
chrI	809310	809320	11	+	CRISPR_repeat	No	-	-	-	-
chrI	1302840	1302850	11	+	CRISPR_repeat	No	-	-	-	-
chrI	1542856	1542866	11	+	CRISPR_repeat	No	-	-	-	-
chrI	1913303	1913313	11	+	CRISPR_repeat	No	-	-	-	-
chrI	3418241	3418251	11	+	CRISPR_repeat	No	-	-	-	-
chrI	60811	60821	11	-	CRISPR_repeat	No	-	-	-	-
chrI	702167	702177	11	-	CRISPR_repeat	No	-	-	-	-
chrI	1100592	1100602	11	-	CRISPR_repeat	No	-	-	-	-
chrI	1654039	1654049	11	-	CRISPR_repeat	No	-	-	-	-
chrI	3078358	3078368	11	-	CRISPR_repeat	No	-	-	-	-
chrI	3199865	3199875	11	-	CRISPR_repeat	No	-	-	-	-
chrI	3456144	3456154	11	-	CRISPR_repeat	No	-	-	-	-
chrI	3534691	3534701	11	-	CRISPR_repeat	No	-	-	-	-
chrI	3590602	3590612	11	-	CRISPR_repeat	No	-	-	-	-
chrI	1500802	1500827	26	-	isrJ_Hfq_binding	No	-	-	-	-
chrI	1690756	1690786	31	+	Deinococcus_Y_RNA	No	-	-	-	-
chrI	2637082	2637128	47	+	EAV_LTH	No	-	-	-	-
chrI	1479823	1479891	69	-	FinP	Yes	-	-	-	-
chrI	1302364	1302469	106	+	GRIK4_3p_UTR	No	-	-	-	-
chrI	261127	261199	73	-	Gurken	No	-	-	-	-
chrI	808656	808703	48	-	Hairpin	No	-	-	-	-
chrI	2640802	2640828	27	+	IRE	No	-	-	-	-
chrI	106391	106435	45	+	K10_TLS	No	-	-	-	-
chrI	270271	270311	41	+	K10_TLS	No	-	-	-	-
chrI	635829	635873	45	-	K10_TLS	No	-	-	-	-
chrI	2342917	2342938	22	+	PSLVbeta_UPD-PK2	No	-	-	-	-
chrI	2673819	2673840	22	+	PSLVbeta_UPD-PK2	No	-	-	-	-
chrI	3241758	3241875	118	+	PyrR	No	-	-	-	-
chrI	2775835	2775965	131	+	PyrR	No	-	-	-	-
chrI	1313428	1313543	116	+	PyrR	Yes	-	-	-0,44	-
chrI	2766380	2766474	95	-	PyrR	No	-	-	-	-
chrI	2167599	2167685	87	-	RtT	No	-	-	-	-
chrI	1445873	1445964	92	-	RtT	No	-	-	-	-
chrI	2822907	2822950	44	+	S-element	No	-	-	-	-
chrI	252065	252124	60	-	S-element	Yes	-	-	-	-
chrI	1658079	1658175	97	-	S15	No	-	-	-	-
chrI	2384773	2384794	22	-	SBRMV1_UPD-PKd	No	-	-	-	-
chrI	1850854	1850921	68	-	SECIS	Yes	-	-	-	-
chrI	1870392	1870470	79	-	SNORA55	No	-	-	-	-
chrI	1334839	1334907	69	+	SNORD15	No	-	-	-	-
chrI	957439	957526	88	-	SNORD21	No	-	-	-	-
chrI	2399635	2399723	89	-	SNORD34	No	-	-	-	-
chrI	2760326	2760398	73	-	SNORD59	No	-	-	-	-
chrI	79478	79550	73	-	SNORD59	No	-	-	-	-
chrI	1397574	1397618	45	+	SNORD70	No	-	-	-	-
chrI	2682080	2682156	77	+	SNORD70	No	-	-	-	-
chrI	2941931	2941996	66	+	SNORD70	No	-	-	-	-

- Continued next page -

Chromosome	Start	End	Length	Strand	Description	Active	Differentially expressed (logFC)			
							BvsP_48	BvsP_120	B_120vs48	P_120vs48
chrI	588261	588310	50	-	SNORD70	No	-	-	-	-
chrI	2273136	2273213	78	-	SNORD70	No	-	-	-	-
chrI	229033	229132	100	+	SNORD86	No	-	-	-	-
chrI	1864077	1864141	65	+	SNORD98	No	-	-	-	-
chrI	336293	336428	136	+	Telomerase-cil	No	-	-	-	-
chrI	1176691	1176823	133	-	Telomerase-cil	No	-	-	-	-
chrI	1201358	1201378	21	-	UPD-PKg	No	-	-	-	-
chrI	1221813	1221905	93	+	bantam	No	-	-	-	-
chrI	2479052	2479120	69	-	ctRNA_pGA1	No	-	-	-	-
chrI	2766386	2766450	65	+	ctRNA_pND324	No	-	-	-	-
chrI	1433265	1433350	86	-	ctRNA_pND324	Yes	-	-	-	-
chrI	3241798	3241883	86	+	ctRNA_pT181	No	-	-	-	-
chrI	906249	906321	73	+	ctRNA_pT181	No	-	-	-	-
chrI	1965107	1965175	69	+	ctRNA_pT181	Yes	-	-	-	-
chrI	2988832	2988905	74	+	ctRNA_pT181	No	-	-	-	-
chrI	982021	982108	88	+	ctRNA_pT181	No	-	-	-	-
chrI	805787	805877	91	+	ctRNA_pT181	No	-	-	-	-
chrI	1512817	1512906	90	-	ctRNA_pT181	Yes	-	-	-	-
chrI	2581761	2581847	87	-	ctRNA_pT181	No	-	-	-	-
chrI	551851	551953	103	-	ctRNA_pT181	No	-	-	-	-
chrI	3324073	3324132	60	+	nos_TCE	No	-	-	-	-
chrI	2465099	2465153	55	+	sR11	No	-	-	-	-
chrI	1495918	1495968	51	+	sR11	No	-	-	-	-
chrI	398673	398733	61	+	sR15	No	-	-	-	-
chrI	3394232	3394298	67	+	sR15	No	-	-	-	-
chrI	1708862	1708926	65	+	sR15	No	-	-	-	-
chrI	2953642	2953699	58	+	sR2	No	-	-	-	-
chrI	1309456	1309505	50	+	sR2	No	-	-	-	-
chrI	63511	63573	63	+	sR21	No	-	-	-	-
chrI	270442	270497	56	+	sR33	Yes	-	-	-	-
chrI	1533334	1533388	55	-	sR48	No	-	-	-	-
chrI	1556386	1556456	71	-	sn2841	No	-	-	-	-
chrI	2398082	2398209	128	+	snR13	No	-	-	-	-
chrI	1237121	1237215	95	+	snR58	Yes	-	-	-	-
chrI	1239485	1239573	89	+	snR62	No	-	-	-	-
chrI	2482774	2482842	69	+	snR62	No	-	-	-	-
chrI	174215	174355	141	-	snoJ26	No	-	-	-	-
chrI	187992	188057	66	-	snoJ26	No	-	-	-	-
chrI	801325	801413	89	+	snoM1	Yes	-	-	-	-
chrI	752727	752774	48	-	snoMe28S-Am982	No	-	-	-	-
chrI	3470065	3470145	81	+	snoR11	Yes	-	-	1,63	-
chrI	145230	145314	85	-	snoR11	No	-	-	-	-
chrI	1709892	1709965	74	+	snoR12	No	-	-	-	-
chrI	2561073	2561149	77	-	snoR160	No	-	-	-	-
chrI	413407	413472	66	-	snoR28	No	-	-	-	-
chrI	479500	479591	92	+	snoR30	No	-	-	-	-
chrI	332258	332294	37	-	snoR31	No	-	-	-	-
chrI	2656210	2656306	97	-	snoR31	No	-	-	-	-
chrI	218105	218191	87	-	snoR43	No	-	-	-	-
chrI	2839461	2839507	47	-	snoR4a	No	-	-	-	-
chrI	2783973	2784014	42	+	snoR53Y	No	-	-	-	-
chrI	1003769	1003806	38	+	snoR53Y	Yes	-	-	-	-
chrI	1831833	1831900	68	+	snoR53Y	No	-	-	-	-
chrI	469141	469207	67	-	snoR53Y	No	-	-	-	-
chrI	365016	365081	66	+	snoR64a	No	-	-	-	-
chrI	2895532	2895594	63	+	snoR72	No	-	-	-	-
chrI	707984	708037	54	+	snoR72	No	-	-	-	-
chrI	427418	427513	96	+	snoR98	Yes	-	-	-1,29	-
chrI	2991310	2991387	78	-	snoR99	No	-	-	-	-
chrI	121064	121120	57	-	snoU35	No	-	-	-	-
chrI	377715	377779	65	-	snoU43C	No	-	-	-	-
chrI	1180607	1180686	80	-	snoU83D	No	-	-	-	-
chrI	1616861	1616909	49	+	snoZ118	No	-	-	-	-
chrI	575330	575431	102	-	snoZ118	No	-	-	-	-
chrI	608866	608946	81	-	snoZ122	No	-	-	-	-
chrI	3168029	3168092	64	+	snoZ13_snr52	No	-	-	-	-
chrI	2409314	2409373	60	+	snoZ13_snr52	Yes	-	-	-0,95	-
chrI	60462	60556	95	-	snoZ13_snr52	No	-	-	-	-
chrI	30984	31055	72	-	snoZ155	No	-	-	-	-
chrI	705749	705803	55	-	snoZ159	No	-	-	-	-
chrI	3041968	3042044	77	-	snoZ159	No	-	-	-	-
chrI	3056776	3056852	77	+	snoZ165	No	-	-	-	-

- Continued next page -

Chromosome	Start	End	Length	Strand	Description	Active	Differentially expressed (logFC)			
							BvsP_48	BvsP_120	B_120vs48	P_120vs48
chrI	1915576	1915656	81	-	snoZ168	No	-	-	-	-
chrI	2759895	2759935	41	+	snoZ169	No	-	-	-	-
chrI	2564043	2564092	50	+	snoZ175	No	-	-	-	-
chrI	2014686	2014726	41	-	snoZ175	No	-	-	-	-
chrI	3438373	3438459	87	+	snoZ182	No	-	-	-	-
chrI	350208	350281	74	+	snoZ196	No	-	-	-	-
chrI	635480	635529	50	-	snoZ196	No	-	-	-	-
chrI	841244	841333	90	+	snoZ223	No	-	-	-	-
chrI	1668509	1668570	62	+	snoZ223	No	-	-	-	-
chrI	2768448	2768530	83	-	snoZ223	No	-	-	-	-
chrI	1445310	1445399	90	-	snoZ247	No	-	-	-	-
chrI	729296	729411	116	-	snoZ5	No	-	-	-	-
chrI	1520269	1520323	55	+	snoZ7	No	-	-	-	-
chrI	3347127	3347200	74	+	suhB	Yes	-	-	-	-
chrI	3220259	3220347	89	-	suhB	Yes	-	-	-	-
chrI	2963232	2963335	104	-	sxy	No	-	-	-	-
chrI	3585895	3585994	100	-	sxy	Yes	-	-	-	-
chrI	2373341	2373453	113	-	TPP	No	-	-	-	-
chrI	1761250	1761315	66	+	RtT	No	-	-	-	-
chrI	1222604	1222952	349	+	RNaseP	Yes	-	-	-	-
chrI	132665	133015	351	-	tmRNA	No	-	-	-	-
chrI	2620243	2620351	109	-	PK-G12rRNA	No	-	-	-	-
chrI	2922549	2922657	109	-	PK-G12rRNA	No	-	-	-	-
chrII	101839	101849	11	+	CRISPR_repeat	No	-	-	-	-
chrII	114356	114366	11	+	CRISPR_repeat	No	-	-	-	-
chrII	73622	73695	74	+	GRIK4_3p_UTR	No	-	-	-	-
chrII	274101	274193	93	+	SL1	Yes	-0,44	0,58	0,43	-0,58
chrII	47848	47910	63	-	SNORD37	No	-	-	-	-
chrII	10667	10765	99	+	Telomerase-cil	No	-	-	-	-
chrII	102287	102318	32	-	UPD-PKib	Yes	-	-	-	-
chrII	179514	179599	86	+	snoZ102_R77	Yes	-	-	-	-
chrII	139385	139459	75	+	snoZ155	Yes	-	-	-	-
chrII	220582	220642	61	-	sxy	Yes	-	-	-	-
p74	23574	23672	99	-	snoJ26	No	-	-	-	-
p74	57812	57872	61	-	snoR53Y	No	-	-	-	-

Supp. Tab. 8.6: Annotation based on manual and structural curation.

Protein	Gene	Annotation
ABZ96159	*LEPBI_I0012*	DNA polymerase III, delta subunit
ABZ96159	*LEPBI_I0012*	DNA polymerase III, delta subunit
ABZ96192	*LEPBI_I0045*	Methyltransferase
ABZ96198	*LEPBI_I0051*	Outer membrane peptidase
ABZ96213	*LEPBI_I0066*	Putative transcriptional repressor
ABZ96245	*LEPBI_I0098*	Surface layer protein
ABZ96246	*LEPBI_I0099*	Phosphodiesterase- biofilm
ABZ96252	*LEPBI_I0105*	UDP-glucose 4-epimerase
ABZ96257	*LEPBI_I0110*	Lipid binding protein
ABZ96259	*LEPBI_I0112*	Beta propeller fold
ABZ96303	*LEPBI_I0158*	Signal recognition particle (ARN)
ABZ96320	*LEPBI_I0175*	Queuosine biosynthesis protein
ABZ96336	*LEPBI_I0191*	Lipid binding protein
ABZ96371	*LEPBI_I0226*	Putative thioesterase
ABZ96391	*LEPBI_I0246*	Alpha-beta hydrolase
ABZ96392	*LEPBI_I0247*	Thioredoxin fold; peroxiredoxin
ABZ96400	*LEPBI_I0255*	Outer membrane protein; cell-WALL attachment
ABZ96405	*LEPBI_I0260*	Sensor histidine kinase
ABZ96420	*LEPBI_I0276*	Ribonuclease-like protein
ABZ96445	*LEPBI_I0301*	alpha-beta-barrel
ABZ96461	*LEPBI_I0317*	tetratricopeptide repeats (TPR) containing protein
ABZ96511	*LEPBI_I0368*	DINB/YFIT-like putative metalloenzyme fold
ABZ96523	*LEPBI_I0381*	Alpha beta topology; metal transport
ABZ96546	*LEPBI_I0404*	Metalloendopeptidase
ABZ96563	*LEPBI_I0421*	Tautomerase/dehalogenase
ABZ96591	*LEPBI_I0452*	Chaperone protein
ABZ96627	*LEPBI_I0489*	Response regulator aspartate phosphatase
ABZ96669	*LEPBI_I0531*	Metal-binding protein
ABZ96675	*LEPBI_I0537*	Permease YjgP/YjgQ family
ABZ96784	*LEPBI_I0651*	Putative periplasmic protease
ABZ96878	*LEPBI_I0746*	Aminoglycoside phosphotransferase
ABZ96899	*LEPBI_I0768*	Transmembrane oligosaccharyl transferase
ABZ96912	*LEPBI_I0782*	Alpha/beta hydrolase family
ABZ96919	*LEPBI_I0789*	Citrate lyase; beta barrel
ABZ96937	*LEPBI_I0809*	Outer membrane lipoprotein
ABZ96983	*LEPBI_I0856*	Multidrug resistance protein outer membrane protein
ABZ96985	*LEPBI_I0858*	Lipase
ABZ96986	*LEPBI_I0859*	Lipase; alpha-beta hydrolase fold
ABZ96987	*LEPBI_I0860*	Lipase
ABZ97010	*LEPBI_I0885*	Lipase chaperone
ABZ97081	*LEPBI_I0957*	DNA repair
ABZ97112	*LEPBI_I0988*	Conserved lipoprotein LPS cell-WALL
ABZ97202	*LEPBI_I1081*	Diguanylate cyclase; biofilm
ABZ97256	*LEPBI_I1139*	flavoprotein- FAD/NADP-binding rossmann fold
ABZ97362	*LEPBI_I1250*	Glutamine cyclotransferase
ABZ97437	*LEPBI_I1327*	Anti sigma factor FlgM
ABZ97464	*LEPBI_I1354*	Transcriptional regulator
ABZ97465	*LEPBI_I1355*	Thioesterase superfamily
ABZ99603	*LEPBI_II0065*	RNA binding protein

- Continued next page -

Protein	Gene	Annotation
ABZ99608	*LEPBI_II0070*	toxin-like protein
ABZ99625	*LEPBI_II0088*	PilZ domain
ABZ98975	*LEPBI_I2906*	DNA double-strand break repair
ABZ99004	*LEPBI_I2935*	Transcriptional regulator
ABZ99033	*LEPBI_I2965*	glycosyl transferase
ABZ99054	*LEPBI_I2987*	DNA-directed RNA polymerase subunit alpha
ABZ99113	*LEPBI_I3047*	Sensor-type histidine kinase
ABZ98443	*LEPBI_I2352*	Sensor histidine kinase
ABZ98462	*LEPBI_I2371*	Plasmid partition protein
ABZ98474	*LEPBI_I2384*	LEMA protein; bromodomain-like fold
ABZ98487	*LEPBI_I2397*	Coiled-coil; cell division
ABZ98603	*LEPBI_I2518*	Phospho-lipase
ABZ98607	*LEPBI_I2523*	TonB protein; beta-hairpin; transporter
ABZ98611	*LEPBI_I2527*	TCS - Sensor Histidine Kinase
ABZ98614	*LEPBI_I2530*	DNA replication and repair
ABZ98615	*LEPBI_I2531*	ParB/Sulfiredoxin fold
ABZ98642	*LEPBI_I2561*	Putative nucleotide-diphospho-sugar transferase
ABZ98673	*LEPBI_I2594*	Outer membrane assembly lipoprotein YFIO
ABZ98749	*LEPBI_I2671*	Chondroitin ABC lyase
ABZ98771	*LEPBI_I2693*	Putative RNA polymerase
ABZ98783	*LEPBI_I2705*	Carboxyl methyltransferase; membrane protein
ABZ98788	*LEPBI_I2710*	Histone fold protein
ABZ98895	*LEPBI_I2820*	Transcriptional regulator
ABZ98899	*LEPBI_I2824*	Choline-binding protein
ABZ98900	*LEPBI_I2825*	Periplasmic/cell wall glycoside hydrolase
ABZ98902	*LEPBI_I2828*	Antibiotic resistance
ABZ98905	*LEPBI_I2834*	Acetyltransferase
ABZ98945	*LEPBI_I2876*	Diguanylate cyclase; zinc sensor; biofilm
ABZ99214	*LEPBI_I3149*	Outer membrane; OMPA-like fold cell-WALL attachment
ABZ99279	*LEPBI_I3214*	Chaperone
ABZ99355	*LEPBI_I3290*	Intramembrane protease
ABZ99363	*LEPBI_I3298*	Outer membrane phosphate-porin
ABZ99399	*LEPBI_I3335*	Zinc peptidase; alpha/beta barrel
ABZ99412	*LEPBI_I3348*	Pyrrolo-quinoline quinone
ABZ99413	*LEPBI_I3349*	Sensor histidine kinase
ABZ99423	*LEPBI_I3359*	Ankyrin repeat family protein
ABZ99425	*LEPBI_I3361*	Ankyrin repeat family protein
ABZ99435	*LEPBI_I3371*	Glutathionylspermidine synthase
ABZ99483	*LEPBI_I3422*	Transcriptional regulator; biofilm
ABZ99514	*LEPBI_I3454*	DTDP sugar isomerase
ABZ99552	*LEPBI_II0014*	Membrane protein
ABZ99561	*LEPBI_II0023*	Probable protease HTPX homolog; heat shock protein
ABZ99570	*LEPBI_II0032*	Putative signal transduction protein
ABZ99573	*LEPBI_II0035*	ATP-NAD kinase
ABZ99577	*LEPBI_II0039*	TETR-family transcriptional regulator
ABZ99578	*LEPBI_II0040*	Diacylglycerol kinase
ABZ99636	*LEPBI_II0100*	Multidrug transporter
ABZ99637	*LEPBI_II0101*	RNA polymerase sigma factor
ABZ99645	*LEPBI_II0109*	Probable metalloproteinase

- Continued next page -

Protein	Gene	Annotation
ABZ99673	*LEPBI_II0138*	Viral-like DNA integrase
ABZ99676	*LEPBI_II0141*	Membrane protein; signal transduction
ABZ97541	*LEPBI_I1433*	Phospholipase C; membrane/calcium binding
ABZ97630	*LEPBI_I1523*	RNA binding protein
ABZ97791	*LEPBI_I1684*	PilO protein
ABZ97830	*LEPBI_I1724*	Alpha/beta hydrolase fold - probable esterase
ABZ97855	*LEPBI_I1749*	Alpha-beta sandwich; hydrolase
ABZ97888	*LEPBI_I1782*	Lojap-like protein
ABZ97911	*LEPBI_I1805*	Probable two-component response regulator
ABZ97925	*LEPBI_I1819*	AdoMet dependent methyltransferase
ABZ97928	*LEPBI_I1822*	Cell-binding factor 2; SURA-like, chaperone
ABZ97936	*LEPBI_I1830*	lipid binding protein
ABZ97950	*LEPBI_I1844*	Putative ABC transporter permease
ABZ98095	*LEPBI_I1993*	Transmembrane oligosaccharyl transferase
ABZ98172	*LEPBI_I2070*	Inner membrane Glycoside hydrolase family 9
ABZ98229	*LEPBI_I2127*	Alpha/beta hydrolase
ABZ98239	*LEPBI_I2137*	SAM-dependent methyltransferase
ABZ98276	*LEPBI_I2177*	Probable surface protein
ABZ98285	*LEPBI_I2186*	Intramembrane protease
ABZ98312	*LEPBI_I2214*	probable outer membrane protein
ABZ98335	*LEPBI_I2239*	Transcriptional repressor
ABZ98404	*LEPBI_I2309*	periplasmic antiviral protein
ABZ98407	*LEPBI_I2312*	Metal ION transporter; CBS domain
ABZ99684	*LEPBI_II0149*	Phospholipase/carboxylesterase
ABZ99695	*LEPBI_II0160*	Methyl-accepting chemotaxis protein
ABZ99709	*LEPBI_II0174*	Transmembrane Cation efflux system protein
ABZ99725	*LEPBI_II0192*	Glyoxalase/bleomycin resistance protein/dioxygenase
ABZ99743	*LEPBI_II0210*	Carboxyl methyltransferase; membrane protein
ABZ99753	*LEPBI_II0220*	Porin/ outer membrane
ABZ99759	*LEPBI_II0226*	Cell WALL hydrolase
ABZ99768	*LEPBI_II0235*	Sulfate permease family protein
ABZ99781	*LEPBI_II0248*	Chaperone
ABZ99783	*LEPBI_II0250*	Putative lipoprotein
ABZ99795	*LEPBI_II0262*	sialic acid metabolism- kelch repeat- beta-propeller
ABZ99796	*LEPBI_II0263*	Sialic acid metabolism; kelch repeat
ABZ99804	*LEPBI_II0271*	Nitrogen regulatory protein P-II
ABZ99809	*LEPBI_II0276*	Lipase/acylhydrolase
ABZ99820	*LEPBI_p0006*	Putative thioesterase
ABZ99823	*LEPBI_p0009*	Alpha-beta protein
ABZ99839	*LEPBI_p0025*	Pyrrolo-quinoline quinone
ABZ99844	*LEPBI_p0030*	Cysteine peptidase
ABZ99846	*LEPBI_p0032*	Catabolism of external DNA
ABZ99854	*LEPBI_p0042*	Alpha-beta protein
ABZ99857	*LEPBI_p0045*	Cation efflux system protein
ABZ99861	*LEPBI_p0049*	Prevent HOST death protein
ABZ99862	*LEPBI_p0050*	Putative ribonuclease

Conclusión del Capítulo 7

La leptospirosis es una infección causada por diferentes especies del género *Leptospira*. Es una enfermedad endémica que afecta a humanos y animales de producción. Recientemente, se describió la capacidad formadora de biofilms en diferentes cepas de *Leptospira*. Esta capacidad es un reconocido factor de virulencia en muchas especies bacterianas, de las cuales se conocen casi en su mayoría los mecanismos genéticos involucrados en el desarrollo de este fenotipo. Sin embargo, en *Leptospira* la identidad y función de los genes involucrados en la formación de biofilms es prácticamente desconocida.

Haciendo uso de herramientas de genómica comparativa y de la gran cantidad de información genómica que recientemente ha sido generada para diversas cepas de *Leptospira*, se determinó el conjunto de genes asociados a la formación de biofilms. Además, mediante secuenciación masiva del transcriptoma, determinamos el perfil de expresión de genes en diferentes etapas de formación de biofilms. Esta aproximación permitió obtener información acerca de la funcionalidad de los genes responsables de la generación de este fenotipo y encontrar nuevos genes efectores involucrados. Finalmente, la información obtenida fue utilizada para reconstruír las vías metabólicas más importantes relacionadas a la formación de biofilms en *Leptospira*, proporcionando información acerca de la evolución de los determinantes genéticos de este factor de virulencia. Estos resultados son gran impacto en el conocimiento de las bases genéticas de la leptospirosis, el cual en un futuro podrá ser utilizado para mejorar los métodos terapéuticos y planes de contingencia de la enfermedad.

9 Genome Announcements

9.1 *Campylobacter fetus venerealis* biovar. intermedius

Citation:

Iraola G*, Pérez R, Naya H, Paolicchi F, Harris D, Lawley TD, Rego N, Hernández M, Calleros L, Carretto L, Velilla A, Morsella A, Méndez A, Gioffre A. (2013) **Complete genome sequence of *Campylobacter fetus* subsp. *venerealis* biovar. intermedius, isolated from the prepuce of a bull**. *Genome Announcements*. 1(4):e00526-13.

* Corresponding author

9.1.1 Announcement. *Campylobacter fetus* subsp. *venerealis* is the causative agent of bovine genital campylobacteriosis, a sexually transmitted disease distributed worldwide. *Campylobacter fetus* subsp. *venerealis* biovar Intermedius strains differ in their biochemical behavior and are prevalent in some countries. We report the first genome sequence for this biovar, isolated from bull prepuce.

Campylobacter fetus is an important veterinary pathogen. This species is currently divided into two subspecies, *Campylobacter fetus* subsp. *fetus*, causative of abortion in sheep, and *Campylobacter fetus* subsp. *venerealis*, the etiologic agent of bovine genital campylobacteriosis [245], a disease that has spread worldwide and causes economic losses mainly in countries where natural breeding is frequent [139]. A distinct group of *C. fetus* strains known as *C. fetus* subsp. *venerealis* biovar Intermedius has also been determined; these strains phenotypically resemble *C. fetus* subsp. *venerealis*, but they react positively to the H_2S test (typically positive for *C. fetus* subsp. *fetus*) [138]. In recent years, an increase in the prevalence of this biovar has been noticed in some countries (e.g., South Africa) [313]. However, the lack of genomic information for these atypical strains has hindered the development of molecular diagnostic tools and the study of their genomic evolution. Here we present the first complete genome sequence for *Campylobacter fetus* subsp. *venerealis* biovar intermedius INTA 99/541, isolated from the prepuce of a

naturally infected bull.

Sequencing was performed on an Illumina Hi-Seq 2000 platform and generated 13,953,630 paired-end reads (2×100 cycles). The resulting library was first corrected using ALLPATHS-LG [148] and then assembled with Velvet software [149], producing 111 contains with an average coverage of 130-fold. The assembly quality was improved using PAGIT toolkit [150], based on the genome sequence of *C. fetus* subsp. *fetus* 82-40 (accession no. NC_008599) as the reference. The final assembly quality was evaluated with an assembly likelihood estimator (ALE) [151]. The resulting pseudo molecule produced by contig scaffolding was automatically annotated with RAST [152].

Campylobacter fetus subsp. *venerealis* biovar Intermedius INTA 99/541 has a circular chromosome of 1,774,509 bp with an average GC content of 33%, including 2,421 putative protein-coding open reading frames (1,36$genesperkb$), 3 rRNA operons, and 40 tRNA genes. BLAST analysis between those contigs that were not used in chromosome scaffolding and the GenBank plasmids database revealed high sequence homology with the pTet *Campylobacter jejuni* plasmid (accession no. NC_008790) and strong synteny conservation of Cpp protein-coding genes, important for plasmid mobilization [175]. However, more extensive analyses are needed to confirm that this strain is carrier of extrachromosomal replicons.

Comparison between *C. fetus* subsp. *venerealis* biovar Intermedius INTA 99/541, *C. fetus* subsp. *venerealis* Azul-94 [175], *C. fetus* subsp. *venerealis* NTCT 10354^T [314], and *C. fetus* subsp. *fetus* 82-40 genomes using the Artemis comparison tool [315] revealed that sequence identity and synteny are conserved along genomes. Further analysis of these genomes will provide information regarding the basis of the unique physiological and biochemical features of *C. fetus* subsp. *venerealis* biovar intermedius. Moreover, he availability of the first genome from this organism is an important achievement in the development of specific molecular tools for diagnosis and will shed light on the genomic evolution of *Campylobacter* species, although a representative number of genomes for this biovar will be needed to conduct more robust comparisons.

Nucleotide sequence accession numbers. This whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession number ASTK00000000. The version described in this paper is version ASTK01000000.

9.1.2 Acknowledgements. This work was funded by Wellcome Trust grant 098051, Comisión Sectorial de Investigaci'on Científica (CSIC), and INTA Argentina Projects PE 242121 and PNSA-1115053. We thank Gordon Dougan and Julian Parkhill for their useful advice during this work.

9.1.3 References

[138] M Veron, R Chatelain, *International Journal of Systematic and Evolutionary Microbiology* **1973**, *23*, 122–134.

[139] G. Mshelia, J. Amin, Z Woldehiwet, R. Murray, G. Egwu, *Reproduction in Domestic Animals* **2010**, *45*, e221–e230.

[148] J. Butler, I. MacCallum, M. Kleber, I. A. Shlyakhter, M. K. Belmonte, E. S. Lander, C. Nusbaum, D. B. Jaffe, *Genome research* **2008**, *18*, 810–820.

[149] D. R. Zerbino, E. Birney, *Genome Res.* **2008**, *18*, 821–829.

[150] M. T. Swain, I. J. Tsai, S. A. Assefa, C. Newbold, M. Berriman, T. D. Otto, *Nat Protoc* **2012**, *7*, 1260–1284.

[151] S. C. Clark, R. Egan, P. I. Frazier, Z. Wang, *Bioinformatics* **2013**, bts723.

[152] R. K. Aziz, D. Bartels, A. A. Best, M. DeJongh, T. Disz, R. A. Edwards, K. Formsma, S. Gerdes, E. M. Glass, M. Kubal, F. Meyer, G. J. Olsen, R. Olson, A. L. Osterman, R. A. Overbeek, L. K. McNeil, D. Paarmann, T. Paczian, B. Parrello, G. D. Pusch, C. Reich, R. Stevens, O. Vassieva, V. Vonstein, A. Wilke, O. Zagnitko, *BMC Genomics* **2008**, *9*, 75.

[153] L. Li, C. J. Stoeckert, D. S. Roos, *Genome research* **2003**, *13*, 2178–2189.

[175] P. M. Moolhuijzen, A. E. Lew-Tabor, B. M. Wlodek, F. G. Agüero, D. J. Comerci, R. A. Ugalde, D. O. Sanchez, R. Appels, M. Bellgard, *BMC microbiology* **2009**, *9*, 86.

[245] L. Debruyne, D. Gevers, P. Vandamme, **2008**.

[313] T Schmidt, E. H. Venter, J. Picard, *Journal of the South African Veterinary Association* **2010**, *81*, 87–92.

[314] A. P. R. Stynen, A. P. Lage, R. J. Moore, A. M. Rezende, P. de Cássia Ruy, N. Daher, D. de Melo Resende, S. S. de Almeida, S. de Castro Soares, V. A. C. de Abreu, et al., *Journal of bacteriology* **2011**, *193*, 5871–5872.

[315] T. J. Carver, K. M. Rutherford, M. Berriman, M.-A. Rajandream, B. G. Barrell, J. Parkhill, *Bioinformatics* **2005**, *21*, 3422–3423.

[316] C. Coitinho, G. Greif, C. Robello, P. Laserra, E. Willery, P. Supply, *European Respiratory Journal* **2014**, *43*, 903–906.

[317] C. Allix-Béguec, P. Supply, M. Wanlin, P. Bifani, M. Fauville-Dufaux, *European Respiratory Journal* **2008**, *31*, 1077–1084.

[318] H. Li, R. Durbin, *Bioinformatics* **2009**, *25*, 1754–1760.

[319] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, et al., *Genome research* **2010**, *20*, 1297–1303.

9.2 A rapidly-progressing tuberculosis in Montevideo

Citation:

Greif G, Iraola G, Berná L, Coitinho C, Rivas CM, Naya H, Robello C*. (2014) **Complete genome sequence of *Mycobacterium tuberculosis* MtURU-001, isolated from a rapidly progressing outbreak in Uruguay**. *Genome Announcements*. 2(1):e01220-13.

* Corresponding author

9.2.1 Announcement. Despite efficient control programs, large clonal outbreaks of tuberculosis (TB) may arise in low-risk populations. Recently, an unusual TB outbreak was reported in Uruguay, reaching an elevated disease attack rate (53 to 69%). Here, we report the genome sequence of the *Mycobacterium tuberculosis* strain associated with this rapidly progressing outbreak, named MtURU-001.

Recently, we reported an unusual tuberculosis (TB) outbreak centered on a professional basketball team in Montevideo, Uruguay, a country with a low TB incidence [316]. In August 2008, a young male member of the basketball team was diagnosed with TB, and a chest X-ray indicated a bilateral pulmonary form with cavities. TB was bacteriologically confirmed 20 days later and the patient was cured after first-line treatment. As described in Coitinho *et al.* [316], following this index case, six other team members who lived at the same place for a week and four other contacts were successively diagnosed with TB over the next 2.5 years. All patients (ranging between 17 and 23 years of age) were immunocompetent, athletic, and wealthy. No other comorbidity factors were detected.

Despite control programs, large clonal TB outbreaks can develop even in low-incidence countries, reflecting ongoing disease transmission [317]. The Mycobacterium tuberculosis strain showed an elevated disease attack rate (53 to 69%) that sharply contrasts with the lifetime risk of developing active TB, being estimated at 10% among infected individuals in the general population. We report here the draft genome sequence of the TB isolate from the index case.

Sequencing was performed at the Institut Pasteur de Montevideo on an Illumina Genome Analyzer IIx platform and generated $2,379,897$ paired-end reads (2×100 cycles). The resulting library was first corrected using ALLPATHS-LG [148] and then assembled with Velvet software [149], producing 195 contigs with an average coverage of 84-fold. The assembly quality was improved using the PAGIT toolkit [150], based on the genome sequence of *M. tuberculosis* H37Rv (GenBank accession no. NC_000962) as a reference strain.

The final assembly quality was evaluated with the Assembly Likelihood Estimator (ALE) software [151], and the assembly was automatically annotated with RAST [152].

M. tuberculosis MtURU-001 has a circular chromosome of 4,378,296 bp, with an average GC content of 65%, including 4,314 protein-encoding genes, 1 rRNA operon, and 45 tRNA genes. In comparison with *M. tuberculosis* H37Rv, 4,096 orthologous groups were defined with OrthoMCL [153] and 1,016 polymorphisms were identified using the Burrows-Wheeler Aligner (BWA) [318] and GATK [319]. A subset of 849 polymorphisms (802 single-nucleotide polymorphisms and 47 indels) were inside coding sequences, and 480 affect protein sequences, especially 24 that introduced stop codons disrupting several hypothetical proteins, one transcriptional regulator, 2 genes for the haloacid dehalogenase (HAD) superfamily, and 3 involved in lipid metabolism. Further comparative genomics across this genome may provide genotype-phenotype associations that might explain the rapid progression of this unusual outbreak.

9.2.2 Acknowledgements. This work was funded by the Comisión Honoraria de Lucha Antituberculosa y Enfermedades Prevalentes (MSP, Uruguay) and the Institut Pasteur de Montevideo and by a fellowship from INNOVA-Uruguay (to LB).

9.2.3 References

[138] M Veron, R Chatelain, *International Journal of Systematic and Evolutionary Microbiology* **1973**, *23*, 122–134.

[139] G. Mshelia, J. Amin, Z Woldehiwet, R. Murray, G. Egwu, *Reproduction in Domestic Animals* **2010**, *45*, e221–e230.

[148] J. Butler, I. MacCallum, M. Kleber, I. A. Shlyakhter, M. K. Belmonte, E. S. Lander, C. Nusbaum, D. B. Jaffe, *Genome research* **2008**, *18*, 810–820.

[149] D. R. Zerbino, E. Birney, *Genome Res.* **2008**, *18*, 821–829.

[150] M. T. Swain, I. J. Tsai, S. A. Assefa, C. Newbold, M. Berriman, T. D. Otto, *Nat Protoc* **2012**, *7*, 1260–1284.

[151] S. C. Clark, R. Egan, P. I. Frazier, Z. Wang, *Bioinformatics* **2013**, bts723.

[152] R. K. Aziz, D. Bartels, A. A. Best, M. DeJongh, T. Disz, R. A. Edwards, K. Formsma, S. Gerdes, E. M. Glass, M. Kubal, F. Meyer, G. J. Olsen, R. Olson, A. L. Osterman, R. A. Overbeek, L. K. McNeil, D. Paarmann, T. Paczian, B. Parrello, G. D. Pusch, C. Reich, R. Stevens, O. Vassieva, V. Vonstein, A. Wilke, O. Zagnitko, *BMC Genomics* **2008**, *9*, 75.

[153] L. Li, C. J. Stoeckert, D. S. Roos, *Genome research* **2003**, *13*, 2178–2189.

[175] P. M. Moolhuijzen, A. E. Lew-Tabor, B. M. Wlodek, F. G. Agüero, D. J. Comerci, R. A. Ugalde, D. O. Sanchez, R. Appels, M. Bellgard, *BMC microbiology* **2009**, *9*, 86.

[245] L. Debruyne, D. Gevers, P. Vandamme, **2008**.

[313] T Schmidt, E. H. Venter, J. Picard, *Journal of the South African Veterinary Association* **2010**, *81*, 87–92.

[314] A. P. R. Stynen, A. P. Lage, R. J. Moore, A. M. Rezende, P. de Cássia Ruy, N. Daher, D. de Melo Resende, S. S. de Almeida, S. de Castro Soares, V. A. C. de Abreu, et al., *Journal of bacteriology* **2011**, *193*, 5871–5872.

[315] T. J. Carver, K. M. Rutherford, M. Berriman, M.-A. Rajandream, B. G. Barrell, J. Parkhill, *Bioinformatics* **2005**, *21*, 3422–3423.

[316] C. Coitinho, G. Greif, C. Robello, P. Laserra, E. Willery, P. Supply, *European Respiratory Journal* **2014**, *43*, 903–906.

[317] C. Allix-Béguec, P. Supply, M. Wanlin, P. Bifani, M. Fauville-Dufaux, *European Respiratory Journal* **2008**, *31*, 1077–1084.

[318] H. Li, R. Durbin, *Bioinformatics* **2009**, *25*, 1754–1760.

[319] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, et al., *Genome research* **2010**, *20*, 1297–1303.

9.3 An isoniazid-resistant tuberculosis isolate

Citation:

Berná L, Iraola G, Greif G, Coitinho C, Rivas CM, Naya H, Robello C*. (2014) **Whole-genome sequencing of an isoniazid-resistant clinical isolate of *Mycobacterium tuberculosis* strain MtURU-002 from Uruguay**. *Genome Announcements*. 2(4):e00655-14.

* Corresponding author

9.3.1 Announcement. The incidence of tuberculosis in Uruguay has been effectively reduced to <30 per 100,000 population, although an increase in non-risk populations in the last few years is evident. Here, we present the genome sequence of *Mycobacterium tuberculosis* strain MtURU-002 isolated from a patient showing bilateral pulmonary tuberculosis that was resistant to isoniazid.

Tuberculosis (TB), caused by infection with Mycobacterium tuberculosis, constitutes a major cause of morbidity and mortality worldwide, ranking as the second leading cause of death from a single infectious agent, after human immunodeficiency virus. In Uruguay, the National Tuberculosis Program has effectively reduced the incidence of TB to 30 per 100,000 population, with 600 to 700 new cases per year (data available at the World Health Organization Web page (`http://www.who.int`). However, in the last 5 years, there has been an increase in TB incidence not only in high-risk populations (patients with human immunodeficiency virus and TB co-infection, those in poverty, and prisoners) but also in non-risk populations, such as in our recent report of 11 cases of well-nourished young subjects affected by the disease [316]. In this context, the whole-genome study of isolates from different populations (high- and low-risk) becomes a necessity in order to perform future comparative genomic studies and to determine new molecular markers of pathogenicity and transmissibility, among other aims. In this work, we performed the complete genome sequencing of a clinical isolate from a patient showing bilateral pulmonary tuberculosis that was resistant to isoniazid.

Sequencing was performed at the Institut Pasteur de Montevideo on an Illumina platform. A total of 1,496,856 paired-end reads (2×100 cycles) were generated; the reads were corrected using ALLPATHS-LG [148], and then Velvet software [149] was used for the *de novo* assembly. A total of 169 contigs were found, with an average coverage of 69-fold. Using the reference genome of *M. tuberculosis* H37Rv (accession no. *NC*_000962), the assembly quality

was further improved through the PAGIT toolkit [150] and evaluated with the Assembly Likelihood Estimator (ALE) software [151]. Finally, automatic annotation was performed using RAST [152]. *M. tuberculosis* strain MtURU-002 has a total of 4,324,103 bp, with an average GC content of 63%. It contains 4,328 predicted coding sequences (CDSs), 1 rRNA operon, and 45 tRNA genes. Single nucleotide polymorphisms versus *M. tuberculosis* H37Rv were identified using BWA [318] and the GATK pipeline [319]. A total of 540 single nucleotide polymorphisms and 35 indels were found. Of them, 482 belong to coding sequences, and 8 introduce stop codons disrupting membrane protein-related genes involved in lipid metabolism or cell wall processes.

As mentioned, drug sensitivity analysis revealed the isolate to be resistant to isoniazid. Remarkably, we did not identify any reported mutation related to this resistance (`https://tbdreamdb.ki.se`). However, we found a novel mutation, G471S, in the *iniB* gene (isoniazid inducible gene) that might explain resistance. Further studies should be done to evaluate the ability of the iniB gene to confer isoniazid resistance.

Nucleotide sequence accession number. This whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession no. JNGE00000000.

9.3.2 Acknowledgements. This work was funded by the Comisión Honoraria de Lucha Antituberculosa y Enfermedades Prevalentes (MSP, Uruguay), Agencia Nacional de Investigación e Innovación (Uruguay) grant DCI-ALA/2011/023-502 "Contrato de apoyo a las políticas de innovación y cohesión territorial", and FOCEM (MERCOSUR Structural Convergence Fund), COF 03/11.

9.3.3 References

[138] M Veron, R Chatelain, *International Journal of Systematic and Evolutionary Microbiology* **1973**, *23*, 122–134.

[139] G. Mshelia, J. Amin, Z Woldehiwet, R. Murray, G. Egwu, *Reproduction in Domestic Animals* **2010**, *45*, e221–e230.

[148] J. Butler, I. MacCallum, M. Kleber, I. A. Shlyakhter, M. K. Belmonte, E. S. Lander, C. Nusbaum, D. B. Jaffe, *Genome research* **2008**, *18*, 810–820.

[149] D. R. Zerbino, E. Birney, *Genome Res.* **2008**, *18*, 821–829.

[150] M. T. Swain, I. J. Tsai, S. A. Assefa, C. Newbold, M. Berriman, T. D. Otto, *Nat Protoc* **2012**, *7*, 1260–1284.

[151] S. C. Clark, R. Egan, P. I. Frazier, Z. Wang, *Bioinformatics* **2013**, bts723.

[152] R. K. Aziz, D. Bartels, A. A. Best, M. DeJongh, T. Disz, R. A. Edwards, K. Formsma, S. Gerdes, E. M. Glass, M. Kubal, F. Meyer, G. J. Olsen, R. Olson, A. L. Osterman, R. A. Overbeek, L. K. McNeil, D. Paarmann, T. Paczian, B. Parrello, G. D. Pusch, C. Reich, R. Stevens, O. Vassieva, V. Vonstein, A. Wilke, O. Zagnitko, *BMC Genomics* **2008**, *9*, 75.

[153] L. Li, C. J. Stoeckert, D. S. Roos, *Genome research* **2003**, *13*, 2178–2189.

[175] P. M. Moolhuijzen, A. E. Lew-Tabor, B. M. Wlodek, F. G. Agüero, D. J. Comerci, R. A. Ugalde, D. O. Sanchez, R. Appels, M. Bellgard, *BMC microbiology* **2009**, *9*, 86.

[245] L. Debruyne, D. Gevers, P. Vandamme, **2008**.

[313] T Schmidt, E. H. Venter, J. Picard, *Journal of the South African Veterinary Association* **2010**, *81*, 87–92.

[314] A. P. R. Stynen, A. P. Lage, R. J. Moore, A. M. Rezende, P. de Cássia Ruy, N. Daher, D. de Melo Resende, S. S. de Almeida, S. de Castro Soares, V. A. C. de Abreu, et al., *Journal of bacteriology* **2011**, *193*, 5871–5872.

[315] T. J. Carver, K. M. Rutherford, M. Berriman, M.-A. Rajandream, B. G. Barrell, J. Parkhill, *Bioinformatics* **2005**, *21*, 3422–3423.

[316] C. Coitinho, G. Greif, C. Robello, P. Laserra, E. Willery, P. Supply, *European Respiratory Journal* **2014**, *43*, 903–906.

[317] C. Allix-Béguec, P. Supply, M. Wanlin, P. Bifani, M. Fauville-Dufaux, *European Respiratory Journal* **2008**, *31*, 1077–1084.

[318] H. Li, R. Durbin, *Bioinformatics* **2009**, *25*, 1754–1760.

[319] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, et al., *Genome research* **2010**, *20*, 1297–1303.

10 A quantitative PCR for *Campylobacter fetus*

Citation:

Iraola G, Pérez R, Betancor L, Marandino A, Morsella C, Méndez A, Paolicchi F, Piccirillo A, Tomás G, Velilla A, Calleros L*. (2016) **A novel real-time PCR assay for quantitative detection of *Campylobacter fetus* based on ribosomal sequences**. *BMC Veterinary Research*. Under review.

* Corresponding author

10.1 Abstract

***Campylobacter fetus* is of major concern for animal and human health. Biochemical tests remain as the gold standard for identifying *C. fetus* but operational difficulties and the lack of reproducibility of some tests motivated the development of molecular diagnostic tools. These methods have been successfully tested on bovine isolates but fail to detect some genetically divergent strains isolated from other hosts. The present study describes the development of a highly sensitive real-time PCR assay that targets a unique region of the 16S rRNA gene as a diagnostic tool to identify every *C. fetus* strain. Our assay detected all *C. fetus* tested in this study, including strains that were negative for the assay used as a standard for molecular species identification, but was negative with other *Campylobacter* species. The high performance obtained with our assay supports its usefulness as a fast and cost-effective tool for *C. fetus* identification in routine diagnostics.**

10.2 Introduction

Members of the genus *Campylobacter* are gram-negative epsilonproteobacteria highly adapted to vertebrate hosts. Most species are pathogens of a wide range of livestock species and have extensive reservoirs in wildlife [145, 320].

The species *Campylobacter fetus* shows a remarkable level of intra-specific variation, with three subspecies: *C. fetus* subsp. *fetus* (Cff), *C. fetus* subsp.

venerealis (Cfv), and *C. fetus* subsp. *testudinum* (Cft). Cff and Cfv are classified on the basis of their mechanisms of transmission, clinical presentations and two key biochemical tests (tolerance to glycine and H_2S production) [138]. Cff infects the intestinal tract of several mammalian species and induces abortion in cattle and sheep [138, 321]. In humans, it is an opportunistic microorganism that mainly infects immune-compromised patients [131]. Cfv is a cattle-restricted pathogen with tropism for genital tissues and is the etiological agent of bovine venereal campylobacteriosis, a serious reproductive disease that causes infertility and abortion [139]. Cfv includes a variant, namely Cfv biovar intermedius (Cfvi) that reacts differently to the H_2S test [138]. Cft has been proposed recently to cluster some reptilian and human strains of putative reptilian origin on the basis of notorious genetic divergence from Cff and Cfv [202].

Biochemical tests remain as the gold standard for identifying *C. fetus* and differentiating between Cff and Cfv, but the fastidious growth requirements and the lack of reliability and reproducibility of some biochemical tests [322], due in part to the genetic heterogeneity of some strains, motivated the development of alternative diagnostic methods.

Several studies have endeavored in determining the suitability of different genetic methods for identifying the species *C. fetus* using end-point PCRs. In particular, the multiplex-PCR assay designed by Hum *et al.* [203] has been vastly used for species identification. Diagnosis of *C. fetus* in this assay is achieved using PCR primers that target signature regions of the *cstA* gene, and Cfv identification is based in the *parA* gene. However, genetic divergence in the *cstA* gene could prevent their detection by this assay, as occur in reptilian strains, and thus fails as a general diagnostic tool to identify the species [202].

Other assays for species identification were later designed to target additional genes, like *cpn60*, which encodes the universal 60-kDa chaperonin, and *nahE*, which encodes a sodium/hydrogen exchanger protein [204, 323]. The *cpn60* and *nahE* gene-based methods have been updated to real-time PCR assays using different technologies [324–326]. Both real-time assays have been successfully tested on bovine isolates but may fail to detect some genetically divergent strains, particularly of reptilian origin, which have nucleotide polymorphisms in many genes.

Diagnosis in *C. fetus* can be improved by developing new real-time PCR assays able to detect strains from all subspecies and hosts. These assays should be designed to target highly stable genomic regions that are characteristic for the species. Ribosomal genes are one of the most common DNA regions used to design PCR assays for the identification and detection of microorganisms. The 16S rRNA gene-targeted molecular tools are widely used as its variability has been thoroughly described in all *Campylobacter species* [147, 327–331].

The sequence of the 16S rRNA gene is species-specific within the genus and *C. fetus* has several unique nucleotide markers [332]. Moreover, ribosomal genes are homogenous for *C. fetus* subspecies and have three identical copies per genome allowing a better detection. Despite the obvious advantages of these genes, so far, there is not a real-time PCR assay targeting ribosomal sequences for the specific detection of *C. fetus*.

The present study describes the development of a highly sensitive real-time PCR assay, which targets a unique region of the 16S rRNA gene as a new diagnostic and quantification tool for every *C. fetus* strains.

10.3 Methods

10.3.1 Real-time PCR design. The assay is based on a set of primers that amplifies a 78-bp sequence of the 16S rRNA gene (16SFw: 5'-GCACCTGTCTCAACTTTC-3' and 16SRv: 5'-CCTTACCTGGGCTTGAT-3') and a TaqMan- MGB probe (16SPb: 5'-VIC-ATCTCTAAGAGATTAGTTG-MGB/NFQ-3'), which targets a 19-bp polymorphic region that discriminates strains of *C. fetus* from the remaining *Campylobacter* species and other bacteria. This polymorphic region (Fig. 10.1) was detected by visual inspection of over 3859 partial and complete 16S rRNA gene sequences aligned with T-Coffee [333]. The constructed alignment comprised sequences from all recognized *Campylobacter* species and from unassigned strains belonging to the genus, which were obtained from the SILVA database [158]. BLAST algorithm [79] was used to check *in silico* primers and probe sequence specificity, and to evaluate the occurrence of non-specific matches with the genomes of *C. fetus* and other bacterial species.

10.3.2 Bacterial strains: species and subspecies identification. The real-time PCR assay was tested with a collection of *C. fetus* strains isolated from cattle, humans and reptiles. Two of the strains (INTA 97/C1N3 and INTA 97/608) were assayed also directly from bovine samples of placenta or vaginal mucus, without a previous isolation step. Seven additional strains from four non-*C. fetus* species that occasionally occur in bovine samples were used to verify the specificity of the assays (Tab. 10.1).

Strains were previously typed using bacteriological methods to test the assay specificity. Samples were grown in Brucella semi-solid Broth and *Campylobacter* selective medium under microaerophillic conditions (85% H_2, 5% O_2, 10% CO_2) for 48 h at 37°C. The presumptive *Campylobacter* colonies were tested by catalase and oxidase tests, and grown in Brucella broth (Sigma-Aldrich, St. Louis, USA) with 1%, 1.3%, 1.5% and 1.9% glycine (Sigma-Aldrich), without glycine and in Brucella broth with NaCl and cysteine

(Sigma-Aldrich) to detect H_2S production with a lead acetate paper (Sigma-Aldrich). Sodium selenite reduction test was also performed. Colonies that grew in 1% glycine were classified as *C. fetus fetus* or *C. fetus testudinum* by their positive (Cff) or negative (Cft) H_2S production. Glycine-sensitive colonies were assigned to the subspecies Cfv (H_2S negative) or Cfvi (H_2S positive) (Tab. 10.1). In a total of 60 strains, 25 were Cff, 20 Cfv, 10 Cfvi, 98 one was Cft, and four were not analyzed.

Strains were further characterized using the multiplex-PCRs designed by Hum *et al.* [203] and Iraola *et al.* [86]. Both assays use the same species-specific primers to detect the *cstA* gene and different genes to identify the subspecies. The first method includes a fragment of the *parA* gene as a Cfv marker, and the second uses a fragment of the *virB11* gene (Tab. 10.1).

In cases where multiplex-PCR based methods failed to identify the isolates, molecular identification of species was confirmed by sequencing a fragment of the 16S rRNA gene, which was amplified using the C412F and C1288R primers described by Linton *et al.* [147].

10.3.3 Real-time PCR assays. DNA was extracted from 500 uL of a suspension of live bacteria in a phosphate-buffered saline pH 7.4 solution (1×108 119 CFU/mL), or from 1 mL of preputial washing or vaginal mucus. The QIAamp DNA Mini Kit (Qiagen, Hilden, Germany) was used for all DNA extractions and the DNA purity was measured as the ratio of absorbance at 260 and 280 nm (A260/280) using a Nanodrop 2000 (Thermo Scientific, Waltham, USA).

Real-time PCR was carried out in a 25-μL reaction containing $1 \times TaqMan$ Genotyping Master Mix (Applied Biosystems, Foster City, USA), 1× Custom TaqMan SNP Genotyping Assay (0.9 μM each primer and 0.2 μM probe), and 1 μL genomic DNA. Thermocycling was performed on an ABIPrism 7500 (Applied Biosystems) and consisted of a 5 min incubation step at 50°C, denaturation for 10 min at 95°C, followed by 40 cycles of 15 s at 95°C and 1 min at 60°C, and a final step of 5 min at 70°C. Fluorescence measurements from VIC fluorophore was collected at the 5 min initial incubation stage, at the 60°C step of each cycle, and at the end of the run.

10.3.4 Standard curve generation for analytical testing. To construct the standard curve for the ribosomal probe we generated 10-fold serial dilutions containing 100-107 genome copies/μL. Number of genome copies was determined by the following formula: Y (genome copies/μL)= [X (g/μL) DNA/(nt genome length × 660)] × (6.022 × 135) using the DNA concentration of the dilution (X) and the genome size of the strain Cff 82-40 (1.77 Mb; GenBank accession number NC008599). The log dilution series of *C. fe-*

Table 10.1: Analyzed isolates, discriminated by host, organ, country and year of isolation. Cft: *C. fetus testudinum*. Cff: *C. fetus fetus*. Cfv: *C. fetus venerealis*. Cfvi: *C. fetus venerealis* biovar intermedius. Cf: *C. fetus*. U: unknown. ND: not determined.

Isolate	Host	Organ	Country	Year	Phenotyping[1]	PCR A[2]	PCR B[3]	RT-PCR
RA8/Italy/2011	Turtle	Cloaca	Italy	2011	Cft	No Cf	No Cf	+
A28	Bovine	U	Australia	1978	Cff	Cff	Cff	+
63	Bovine	Prepuce	Uruguay	1980	Cff	Cff	Cff	+
835	Bovine	U	Uruguay	U	Cff	Cfv	Cff	+
F106	Bovine	U	Uruguay	U	Cff	Cff	ND	+
71098	Bovine	Fetal abomasal content	Uruguay	1998	Cff	Cff	Cff	+
C1N3	Bovine	Vaginal mucus	Argentina	1997	Cff	Cff	Cff	+
INTA 04/554	Bovine	Fetal abomasal content	Argentina	2004	Cff	Cff	Cff	+
INTA 90/189	Bovine	Fetal lung	Argentina	1990	Cff	Cfv	Cfv	+
INTA 89/222	Bovine	Prepuce	Argentina	1989	Cff	No Cf/Cfv	No Cf/Cfv	+
INTA 01/165	Bovine	Vaginal mucus	Argentina	2001	Cff	Cff	Cff	+
INTA 12/218	Bovine	Fetal abomasal content	Argentina	2012	Cff	Cfv	Cfv	+
INTA 99/801	Bovine	Prepuce	Argentina	1999	Cff	Cff	Cff	+
INTA 01/064	Bovine	Vaginal mucus	Argentina	2001	Cff	Cff	Cff	+
INTA 04/875	Bovine	Vaginal mucus	Argentina	2004	Cff	Cff	Cff	+
INTA 08/328	Bovine	Fetal lung	Argentina	2008	Cff	Cff	Cff	+
INTA 05/622	Bovine	Fetal abomasal content	Argentina	2005	Cff	Cff	Cfv	+
INTA 11/262	Bovine	Fetal abomasal content	Argentina	2011	Cff	Cfv	Cfv	+
INTA 11/295	Bovine	Fetal abomasal content	Argentina	2011	Cff	Cfv	Cfv	+
INTA 11/685A	Bovine	Vaginal mucus	Argentina	2011	Cff	Cfv	Cff	+
INTA 11/685B	Bovine	Fetal abomasal content	Argentina	2011	Cff	Cfv	Cff	+
INTA 11/677	Bovine	U	Argentina	2011	Cff	Cff	Cff	+
INTA 11/501	Bovine	Vaginal mucus	Argentina	2011	Cff	Cff	Cff	+
INTA 11/408	Bovine	Fetal abomasal content	Argentina	2011	Cff	Cff	Cff	+
INTA 11/356	Bovine	Fetal abomasal content	Argentina	2011	Cff	Cff	Cfv	+
INTA 11/360	Bovine	Fetal lung	Argentina	2011	Cff	Cfv	Cfv	+
NCTC10354T	Bovine	U	England	1962	Cfv	Cff	Cfv	+
D78	Bovine	U	Australia	1978	Cfv	Cfv	Cfv	+
660	Bovine	Fetal abomasal content	Uruguay	2010	Cfv	Cfv	Cfv	+
3726	Bovine	Fetal abomasal content	Uruguay	2010	Cfv	Cfv	Cfv	+
2733	Bovine	Fetal abomasal content	Uruguay	2006	Cfv	Cfv	Cfv	+
2740	Bovine	Fetal abomasal content	Uruguay	2006	Cfv	Cfv	Cfv	+
MCR03	Bovine	Prepuce	Uruguay	2009	Cfv	Cfv	Cfv	+
3837	Bovine	Fetal abomasal content	Uruguay	2010	Cfv	Cfv	Cfv	+
1198	Bovine	U	Uruguay	U	Cfv	Cff	Cfv	+
3598	Bovine	U	Uruguay	U	Cfv	Cff	Cfv	+
2432	Bovine	U	Uruguay	2010	Cfv	Cfv	Cfv	+
2370P	Bovine	Fetal abomasal content	Uruguay	2011	Cfv	Cfv	Cfv	+
2374C	Bovine	Fetal abomasal content	Uruguay	2011	Cfv	Cfv	Cfv	+
27460P	Bovine	Fetal abomasal content	Uruguay	2011	Cfv	Cfv	Cfv	+
INTA 97/608	Bovine	Placenta	Argentina	1997	Cfv	Cfv	Cfv	+
371	Bovine	Vaginal mucus	Argentina	1983	Cfv	Cfv	Cfv	+
INTA 90/264	Bovine	Fetal abomasal content	Argentina	1990	Cfv	Cff	Cfv	+
INTA 05/355	Bovine	Fetal abomasal content	Argentina	2005	Cfv	Cfv	Cfv	+
INTA 95/258	Bovine	Vaginal mucus	Argentina	1995	Cfv	Cff	Cfv	+
INTA 08/382	Bovine	Fetal abomasal content	Argentina	2008	Cfv	Cff	Cfv	+
21	Bovine	U	Australia	1978	Cfvi	ND	ND	+
BL472	Bovine	U	Argentina	1998	Cfvi	Cfv	Cfv	+
INTA 99/541	Bovine	Prepuce	Argentina	1999	Cfvi	Cff	Cfv	+
INTA 97/384	Bovine	Fetal abomasal content	Argentina	1997	Cfvi	Cff	Cfv	+
INTA 98/472	Bovine	Fetal abomasal content	Argentina	1998	Cfvi	Cfv	Cfv	+
INTA 00/305	Bovine	Fetal abomasal content	Argentina	2000	Cfvi	Cff	Cfv	+
INTA 02/146	Bovine	Vaginal mucus	Argentina	2002	Cfvi	Cfv	Cfv	+
INTA 03/596	Bovine	Fetal abomasal content	Argentina	2003	Cfvi	Cff	Cff	+
INTA 07/379	Bovine	Fetal abomasal content	Argentina	2007	Cfvi	Cff	Cfv	+
INTA 06/341	Bovine	Fetal lung	Argentina	2006	Cfvi	Cfv	Cfv	+
H1-UY	Human	Blood	Uruguay	2013	Cf	Cff	Cff	+
HC	Human	Blood	Uruguay	2014	Cf	Cff	Cff	+
70L	Human	Cerebrospinal fluid	Uruguay	2014	Cf	Cff	Cff	+
70H	Human	Blood	Uruguay	2014	Cf	Cff	Cff	+
INTA 08/209	Bovine	Prepuce	Argentina	2008	*C. sputorum*	ND	ND	-
CcHB41	Human	Blood	Uruguay	2014	*C. coli*	ND	ND	-
CjHB32	Human	Blood	Uruguay	2014	*C. jejuni*	ND	ND	-
CjCP3	Chicken	U	Uruguay	2014	*C. jejuni*	ND	ND	-
CcCP60	Chicken	U	Uruguay	2014	*C. coli*	ND	ND	-
Ch99/243	U	U	Argentina	1999	*C. hyointestinalis*	ND	ND	-
NCTC 11562	Pork	U	Inglaterra	1983	*C. hyointestinalis*	ND	ND	-

tus genomes and negative controls containing nuclease-free water were tested with real-time PCR in triplicate and in three independent runs.

Standard curve was generated by plotting threshold cycle (Ct) values per three replicates per standard dilution versus the logarithm of the bacterial genome copies to determine analytical sensitivity and efficiency of the assay. The amplification efficiency was calculated with the Eq. 3, where (k) is the slope of the linear regression line [334, 335]. A value of 1 corresponds to 100% amplification efficiency. The coefficient of determination (R^2) was also assessed and was considered to be suitable when it was higher than 0.980 in a single run [336, 337]. The coefficients of variation (CVs) of Ct values were assessed separately for each standard bacterial dilution by analyzing the replicates of the same analytical run (intra-assay) and the repeated analyses from different analytical runs (inter-assay).

$$E = 10^{-1/k} - 1 \tag{3}$$

10.4 Results

Strains were assigned to *C. fetus* and its subspecies (Cff, Cft, Cfv and Cfvi) using standard bacteriological methods (Tab. 10.1). Additionally, we performed the molecular characterization in the same collection of strains (Tab. 10.1). The results of bacteriological and molecular classification are not always coincident, particularly at the subspecies level. One bovine (INTA 89/222) and the reptilian isolate (RA8/Italy/2011) were phenotypically identified as *C. fetus* but were negative for the *cstA* gene amplicon that functions as marker for *C. fetus*. The bovine isolate was positive for the subspecies (Cfv) markers of both tests and the reptilian isolate was negative. Belonging of these isolates to *C. fetus* was confirmed by sequencing a fragment of the 16SrRNA gene, which unequivocally discriminate between *Campylobacter* species and from other bacterial species [147, 332].

Table 10.2: Intra-and inter-assay reproducibility for the detection of *C. fetus*. CV = coefficient of variation of Ct values [%].

Genome copies/reaction	Intra-assay var. (Ct)	Inter-assay var. (CV)	Mean Ct	CV
1×10^1	-[1]	-	-	-
1×10^2	36,57 - 37,69	0,97 - 2,1	37,13	2,19
1×10^3	33,68 - 34,11	0,48 - 1,15	33,89	1,05
1×10^4	30 - 30,07	0,25 - 0,16	30,03	0,23
1×10^5	26,37 - 26,46	0,14 - 0,27	26,41	0,26
1×10^6	22,62 - 22,86	0,18 - 0,7	22,74	0,73
1×10^7	19,02 - 19,23	0,5 - 0,83	19,12	0,86

[1]: Ct value out of dynamic range.

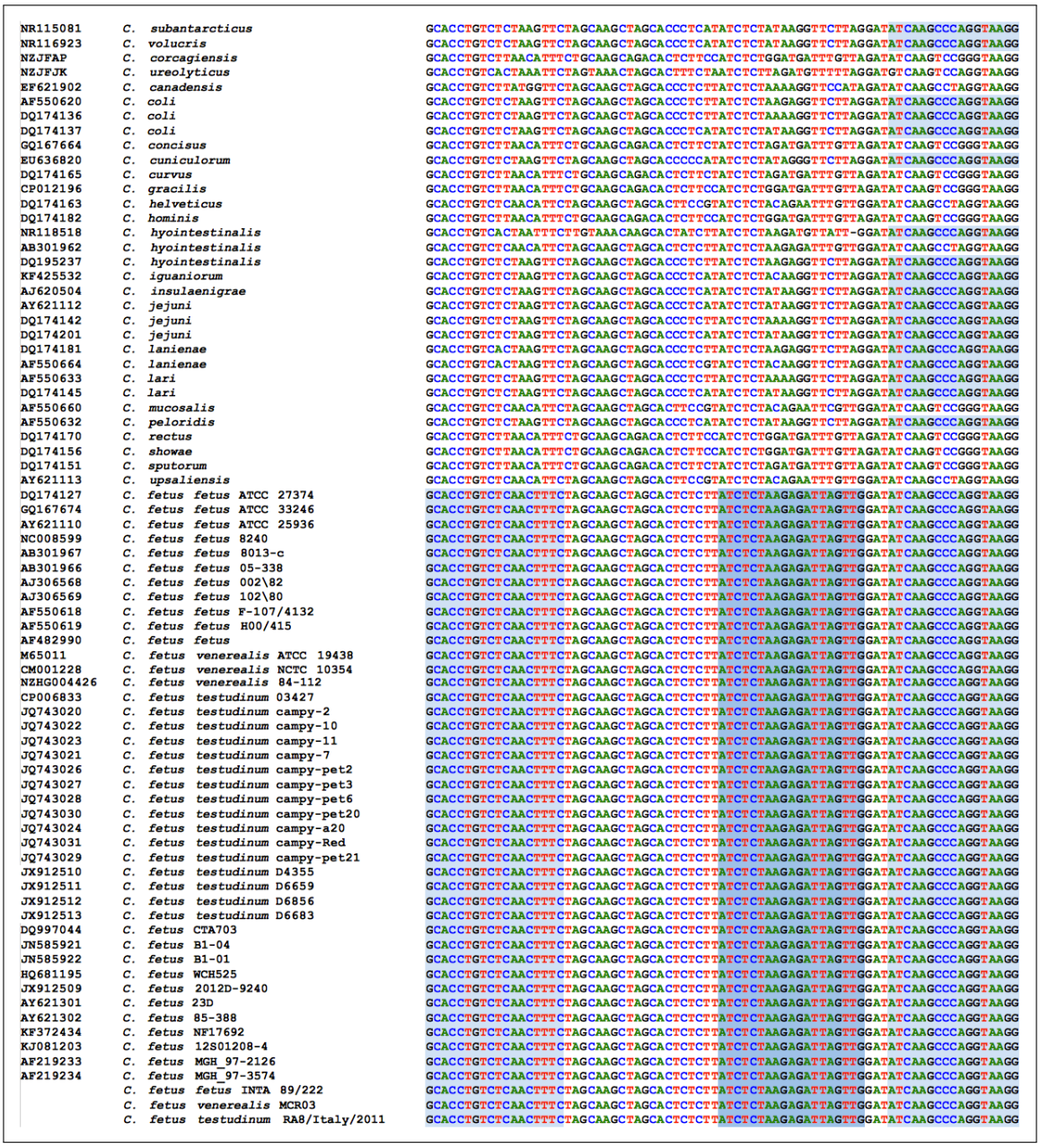


Figure 10.1: Alignment of partial sequences of 16S obtained from databases. Sequences of all species of the genus from which information is available are shown. The sequences of the primers and probe are shaded.

The 16SPb probe is species specific and has a minimum of one (with a

single sequence from *C. hyointestinalis*) and a maximum of nine mismatches with other *Campylobacter* species (e.g. *C. rectus* and *C. showae*). The forward primer's sequence is species specific and has a minimum of one and a maximum of four mismatches with other *Campylobacter* species (Figs. 10.1 and 10.2). The reverse primer's sequence appears identical in some *Campylobacter* species but has one or two differences with others. The combination of primers and probe only matches perfectly with the 16S rRNA gene of *C. fetus* and not with other organisms available in the Genbank database. PCR reactions using template DNA from Cff, Cfv, Cfvi and Cft yielded VIC signals corresponding to *C. fetus*-specific amplification.

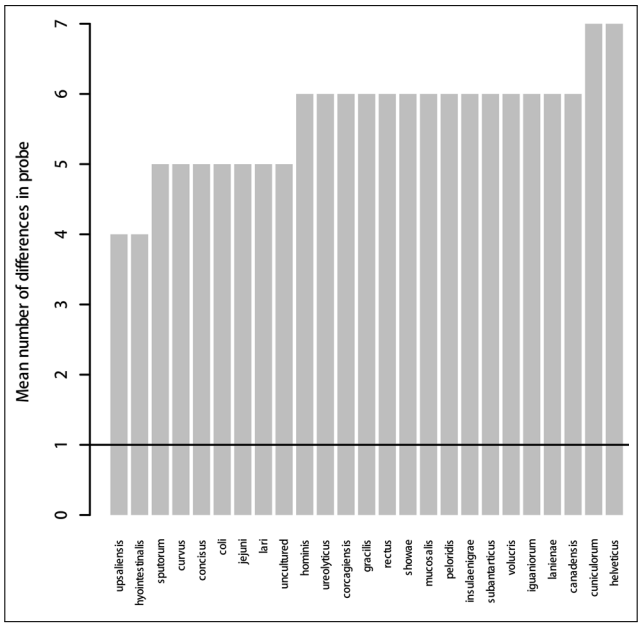


Figure 10.2: Mistmatches in 16SPb. Mean number of differences in probe sequence of non-*C. fetus* species 16S gene.

The analytical testing of the assay was determined using a standard curve (Fig. 10.3). The linear dynamic range of the assay was established between genome copies per reaction. Amplification efficiency and coefficient of determination (R^2) were 93% and 0.9973, respectively. Intra- and inter-assay reproducibility was calculated using coefficient of variation (CV), which showed considerable low values, being the highest 2.19% (Tab. 10.2). No amplification was observed using template DNA from non-*C. fetus* bacterial species used as negative controls (i.e. *C. hyointestinalis*, *C. jejuni*, *C. coli* and *C. sputorum*).

10.5 Discussion

Campylobacter fetus is a pathogen of great relevance for the cattle industry and public health. It is mandatory to diagnose *C. fetus* in cattle to control bovine genital campylobacteriosis. In humans it is necessary to detect this opportunistic pathogen to achieve a better treatment and for epidemiological surveys. Detection of *C. fetus* in humans is difficult because all subspecies are potential pathogens and well-stablished methods would fail to detect strains of reptilian origin [202]. Therefore, the cost-effective, automated and straightforward methods for the unambiguous identification of *C. fetus* are of paramount importance.

Bacteriological analysis, like culture isolation and biochemical tests, are well standardized and extensively used but challenging by the slow growing and few differential phenotypic properties of *C. fetus* [256]. These methods are also laborious and time-consuming, a disadvantage when processing samples at large-scale or delivering a fast diagnosis. To improve the quality and complement the gold-standard bacteriological methods for *C. fetus* detection, some end-point PCR methods have been designed based on the presence of species-specific amplicons [140, 203, 338, 339]; these assays fulfill various criteria such as accuracy, high detection probability and well-standardized protocols for its application and interpretation. Real-time PCR methods have been also designed with the same purpose [323–326] and have provided additional technical improvements to *C. fetus* detection protocols, like the prevention of cross contamination and the minimization of manipulation and running times. However, both end-point and real-time PCR methods described to date have difficulty in dealing with the intra-specific genetic variability of *C. fetus*, failing in capturing all strains from diverse hosts. In comparison to conventional PCR methods, a real-time PCR assay would offer increased sensitivity and an accurate quantification of target DNA to study the dynamics of the bacteria in different hosts and tissues. To the best of our knowledge, there is not a real-time PCR method that uses ribosomal sequences or any other core genome regions for identification of *C. fetus*. Here, we have improved the current molecular methods for *C. fetus* detection by designing a new real-time PCR assay that targets the multi-copy 16S rRNA gene. The variability of these sequences within *Campylobacter* species supports its suitability as a target for amplification-based methods using fluorescent probes. The inclusion in the assay of a TaqMan-MGB probe provides higher specificity, sensitivity and accuracy than traditional TaqMan probes and discriminates between sequences that differ in just one nucleotide [340–342].

Our assay was compared to the *cstA* gene end-point PCR proposed by Hum *et al.* [203] and currently used as standard for molecular diagnosis of *C. fetus*. The bovine sample INTA 89/222 and the reptilian RA8/Italy/2011 could not

be detected by Hum's PCR (Tab. 10.1), revealing that the sensitivity of the method for bovine isolates is not 100% as previously reported [203, 205, 252, 313, 326, 343]. This lack of amplification could be due to the absence of the target *cstA* gene in these strains, or the presence of sequence variations that prevent the correct annealing of primers. Our attempt to amplify a larger region including Hum's region also failed, indicating the absence of this gene in the strain or an even greater sequence divergence (data not shown). To test this hypothesis, it would be necessary to conduct the whole genome analysis of these strains. Sequence variability inside the *cstA* gene is not despicable among different *C. fetus* strains. This idea is supported by the presence of several differences in Hum's primers binding sites in the reptilian Cft 03-427 strain complete genome (GeneBank accession number NC_022759). This explains why the 13 isolates used for description of this subspecies, and the RA8/Italy/2011 strain here analyzed, were negative for Hum's method based on the *cstA* gene [202]. Given the importance of this gene in the metabolism of nitrogen, and the recent discovery of their role in interactions with the host in *C. jejuni* [344], it is necessary to continue investigating its variations and possible roles in *C. fetus*.

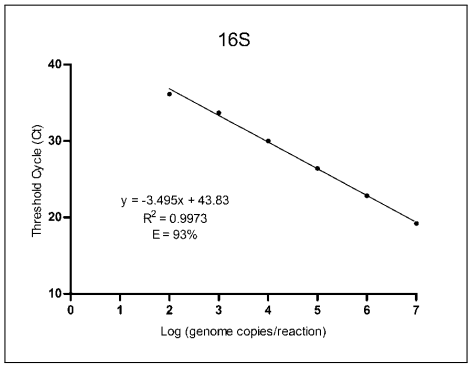


Figure 10.3: Standard curve of developed TaqMan-MGB real-time PCR for *C. fetus* detection. Each point represents the mean Ct of nine different measures (three independent reactions, three replicates each). The curve equation (y), coefficient of determination (R^2) and amplification efficiency (E) are indicated.

Our novel real-time PCR assay detected all *C. fetus* tested in this study, but was negative with other Campylobacter species. The complete identity of primer and probe targets in all *C. fetus* strains deposited in the GenBank database, including reptilian ones, also supports that our assay is expected to detect all subspecies from diverse hosts (Fig. 10.1). These results indicate the excellent sensitivity and specificity of the assay, representing an advantage over *cstA* gene-based methods. In addition, the primers and probe sequences

are conserved in the 16S rRNA gene of the three subspecies, (Fig. 10.1), in contrast with what happens with primers that amplify the *cstA* gene.

The method here described has some advantages over other real-time PCR methods described in the literature. The *nahE* assay reported by van der Graaf-van Bloois *et al.* [326] uses a TaqMan probe that gives high sensibility and detection capability, but the quantification capability of this assay has not been ascertained using a standard curve. It is also uncertain whether this assay would detect reptilian Cft isolates, which were not assayed in the original paper, as the probe and the forward PCR primers have two mismatches each with the Cft reference strain 03-427. The hybridization of primers and probes in the *nahE* gene could be also affected by genomic rearrangements, which are present in the area around this gene in most of the complete genome sequences available in the databases (not shown). The methodology to detect the *cpn60* gene described by Chaban *et al.* [323] uses specific primers and SYBR green chemistry to identify *C. fetus* species, but its performance is suboptimal in samples with low bacterial concentrations [324], such as the uncultivated samples that were successfully tested in the present assay (Tab. 10.1).

In conclusion, the 16S rRNA gene-targeted assay here developed is highly specific and sensitive and constitutes a valuable molecular tool for assessing the presence of *C. fetus*. The method proved to be useful for detecting *C. fetus* in the field, which may help to understand its epidemiological dynamics to implement more specific applications for its control. The high performance obtained with our assay supports its usefulness as a fast and cost-effective tool for *C. fetus* identification in routine diagnostics. For this reasons, this methodology is a good option to establish a new standard in molecular identification of *C. fetus* species.

10.6 Acknowledgemets

LC and GI acknowledge support from the Comisión Sectorial de Investigación Científica (CSIC), and Agencia Nacional de Investigación e Innovación (ANII) fellowship programs from Uruguay.

10.7 References

[79] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **1990**, *215*, 403–410.

[86] G. Iraola, M. Hernandez, L. Calleros, F. Paolicchi, S. Silveyra, A. Velilla, L. Carretto, E. Rodriguez, R. Perez, *J. Vet. Sci.* **2012**, *13*, 371–376.

[131] J. A. Wagenaar, M. A. van Bergen, M. J. Blaser, R. V. Tauxe, D. G. Newell, J. P. van Putten, *Clin. Infect. Dis.* **2014**, *58*, 1579–1586.

[138] M Veron, R Chatelain, *International Journal of Systematic and Evolutionary Microbiology* **1973**, *23*, 122–134.

[139] G. Mshelia, J. Amin, Z Woldehiwet, R. Murray, G. Egwu, *Reproduction in Domestic Animals* **2010**, *45*, e221–e230.

[140] M. A. van Bergen, K. E. Dingle, M. C. Maiden, D. G. Newell, L. van der Graaf-Van Bloois, J. P. van Putten, J. A. Wagenaar, *J. Clin. Microbiol.* **2005**, *43*, 5888–5898.

[145] S. M. Man, *Nat Rev Gastroenterol Hepatol* **2011**, *8*, 669–685.

[147] D. Linton, R. J. Owen, J. Stanley, *Res. Microbiol.* **1996**, *147*, 707–718.

[158] E. Pruesse, C. Quast, K. Knittel, B. M. Fuchs, W. Ludwig, J. Peplies, F. O. Glöckner, *Nucleic acids research* **2007**, *35*, 7188–7196.

[202] C. Fitzgerald, Z. C. Tu, M. Patrick, T. Stiles, A. J. Lawson, M. Santovenia, M. J. Gilbert, M. van Bergen, K. Joyce, J. Pruckler, S. Stroika, B. Duim, W. G. Miller, V. Loparev, J. C. Sinnige, P. I. Fields, R. V. Tauxe, M. J. Blaser, J. A. Wagenaar, *Int. J. Syst. Evol. Microbiol.* **2014**, *64*, 2944–2948.

[203] S. Hum, K. Quinn, J. Brunner, S. L. On, *Aust. Vet. J.* **1997**, *75*, 827–831.

[204] C. Abril, E. M. Vilei, I. Brodard, A. Burnens, J. Frey, R. Miserez, *Clin. Microbiol. Infect.* **2007**, *13*, 993–1000.

[205] L. van der Graaf-van Bloois, W. G. Miller, E. Yee, M. Rijnsburger, J. A. Wagenaar, B. Duim, *J. Clin. Microbiol.* **2014**, *52*, 4183–4188.

[252] K Willoughby, P. Nettleton, M Quirie, M. Maley, G Foster, M Toszeghy, D. Newell, *Journal of applied microbiology* **2005**, *99*, 758–766.

[256] S. On, B Bloch, B Holmes, B Hoste, P Vandamme, *International Journal of Bacteriology* **1996**, *45*, 767–774.

[313] T Schmidt, E. H. Venter, J. Picard, *Journal of the South African Veterinary Association* **2010**, *81*, 87–92.

[320] S. L. On, *Journal of Applied Microbiology* **2001**, *90*.

[321] M. Garcia, G. Ruckerbauer, M. Eaglesome, W. Boisclair, *Canadian Journal of Comparative Medicine* **1983**, *47*, 336.

[322] T. M. Alves, A. P. R. Stynen, K. L. Miranda, A. P. Lage, *Pesq. Vet. Bras* **2011**, *31*, 336–344.

[323] B. Chaban, K. M. Musil, C. G. Himsworth, J. E. Hill, *Applied and environmental microbiology* **2009**, *75*, 3055–3061.

[324] B. Chaban, S. Chu, S. Hendrick, C. Waldner, J. E. Hill, *Canadian Journal of Veterinary Research* **2012**, *76*, 166–173.

[325] A. M. Selim, M. M. Elhaig, W Gaedem, *Veterinaria italiana* **2014**, *50*, 269–275.

[326] L. van der Graaf-van Bloois, M. A. van Bergen, F. J. van der Wal, A. G. de Boer, B. Duim, T. Schmidt, J. A. Wagenaar, *Journal of microbiological methods* **2013**, *95*, 93–97.

[327] K Blom, C. Patton, M. Nicholson, B Swaminathan, *Journal of clinical microbiology* **1995**, *33*, 1360–1362.

[328] P. Cardarelli-Leite, K. Blom, C. M. Patton, M. A. Nicholson, A. G. Steigerwalt, S. B. Hunter, D. J. Brenner, T. J. Barrett, B. Swaminathan, *Journal of clinical microbiology* **1996**, *34*, 62–67.

[329] C. P. Kolbert, D. H. Persing, *Current opinion in microbiology* **1999**, *2*, 299–305.

[330] W. G. Weisburg, S. M. Barns, D. A. Pelletier, D. J. Lane, *Journal of bacteriology* **1991**, *173*, 697–703.

[331] I. Wesley, R. Wesley, M Cardella, F. Dewhirst, B. Paster, *Journal of clinical microbiology* **1991**, *29*, 1812–1817.

[332] G. Gorkiewicz, G. Feierl, C. Schober, F. Dieber, J. Köfer, R. Zechner, E. L. Zechner, *Journal of Clinical Microbiology* **2003**, *41*, 2537–2546.

[333] C. Notredame, D. G. Higgins, J. Heringa, *Journal of molecular biology* **2000**, *302*, 205–217.

[334] D. G. Ginzinger, *Experimental hematology* **2002**, *30*, 503–512.

[335] M. W. Pfaffl, *Nucleic acids research* **2001**, *29*, e45–e45.

[336] M. B. Gašparič, K. Cankar, J. Žel, K. Gruden, *BMC biotechnology* **2008**, *8*, 1.

[337] S. A. Bustin, T. Nolan, *J Biomol Tech* **2004**, *15*, 155–166.

[338] C. Tramuta, D. Lacerenza, S. Zoppi, M. Goria, A. Dondo, E. Ferroglio, P. Nebbia, S. Rosati, *Journal of Veterinary Diagnostic Investigation* **2011**, *23*, 657–664.

[339] G. Wang, C. G. Clark, T. M. Taylor, C. Pucknell, C. Barton, L. Price, D. L. Woodward, F. G. Rodgers, *Journal of Clinical Microbiology* **2002**, *40*, 4744–4747.

[340] I. V. Kutyavin, I. A. Afonina, A. Mills, V. V. Gorn, E. A. Lukhtanov, E. S. Belousov, M. J. Singer, D. K. Walburger, S. G. Lokhov, A. A. Gall, et al., *Nucleic acids research* **2000**, *28*, 655–661.

[341] N. Rousselon, J.-P. Delgenès, J.-J. Godon, *Journal of microbiological methods* **2004**, *59*, 15–22.

[342] R. Alonso, E. Mateo, R. Cisterna, *Journal of microbiological methods* **2007**, *69*, 214–217.

[343] F. Schulze, A. Bagon, W. Müller, H. Hotzel, *Journal of clinical microbiology* **2006**, *44*, 2019–2024.

[344] J. J. Rasmussen, C. S. Vegge, H. Frøkiær, R. Howlett, K. Krogfelt, D. Kelly, H. Ingmer, *Journal of medical microbiology* **2013**, *62*, 1135–1143.

Conclusión

Durante el desarrollo de mis estudios de posgrado tanto a nivel de la Maestría en Bioinformática como del Doctorado en Biología, me he formado en el área de la biología computacional y bioinformática con especial énfasis en el estudio de genomas de microorganismos. Es bien sabido que la genómica ha revolcionado la biología desde finales del siglo XX, y la microbiología puede considerarse una de las disciplinas pioneras en la incorporación de datos genómicos ya que el conocimiento de la información genómica, primero a nivel individual y luego a niveles de poblaciones y comunidades, ha posibilitado refinar un sinfín de metodologías aplicadas a solucionar problemas microbiológicos. En particular, se ha avanzado significativamente en la incorporación de datos genómicos a los protocolos de taxonomía polifásica que han permitido refinar los procedimientos de identificación de nuevos grupos de procariotas. Además, el conocimiento de la información genómica ha sido sumamente relevante en el estudio de microorganismos patógenos, tanto para el diseño de ensayos de diagnóstico molecular, la identificación de factores de virulencia y genes asociados a la resistencia a antimicrobianos, como para el estudio de brotes y su caracterización a nivel global.

En esta Tesis se han desarrollado y aplicado herramientas que permitieron generar conocimiento científico original en varios de los aspectos mencionados anteriormente. En particular, la **parte uno** describe el desarrollo de una herramienta que permite predecir la patogenicidad de una bacteria (para el humano) simplemente a partir del análisis de un conjunto de genes codificados en su genoma. Luego, y a partir de ese conjunto de genes, se describe un modelo básico que permite predecir la emergencia de nuevos patǵenos a partir de eventos de transferencia horizontal en comunidades bacterianas caracterizadas por metagenómica. Por otro lado, se logró el desarrollo de modelos de predicción para los niveles taxonómicos superiores que evidencia una fuerte correlación entre el potencial metabólico de los diferentes grupos y su ubicación en la taxonomía, sustentando la hipótesis de coherencia ecológica. La **parte dos** de la Tesis se focaliza exclusivamente en análisis genómicos realizados en especies del género *Campylobacter*, de gran importancia sanitaria tanto para humanos como para animales de producción. En este contexto, se realizaron aportes generales al conocimiento de los determinantes genómicos para el tropismo de hospedero y la patogenicidad de estas especies, se describió una nueva especie y se realizaron estudios de epidemiología molecular en dos especies de importancia sanitaria para hu-

manos y animales como *C. fetus* y *C. hyointestinalis*. Finalmente, la **parte tres** es en cierto modo un anexo que comprende tres trabajos específicos en tres organismos diferentes: en primer lugar se describe el primer transcriptoma de *Leptospira biflexa* en condiciones de vida libre y formación de biofilms. En segundo lugar, se describen una serie de trabajos focalizados en la generación de genomas completos de cepas de *Mycobacterium tuberculosis* aisladas de casos clínicos en Uruguay, destcándose la presencia de cepas con fenotipos particulares como por ejemplo la multi-resistencia a antibióticos. En tercer lugar, se presenta un trabajo nuevamente con *C. fetus* pero orientado a la búsqueda de marcadores moleculares para su diagnóstico basado en estudios comparativos de sus genes ribosomales.

En resumen, la Tesis abarca diversas temáticas específicas que pueden ser enmarcadas en un neologismo denominado microbiología computacional. Esta combinación de palabras (*computational microbiology*, del inglés) devuelve, aproximadamente, tan solo 25.000 resultados en una búsqueda en Google Scholar limitada al año 2004, pero si la misma búsqueda se realiza desde esa fecha en adelante el número de resultados se mutiplica por más de un orden de magnitud (en tan solo 10 años). Este incremento responde a la irrupción de las tecnologías de secuenciación masiva que trajeron aparejado un desarrollo particularmente vertiginoso de herramientas y aproximaciones analíticas en microbiología. Aún hoy en día, cuando se considera que la era genómica esta instalada e influye directamente en prácticamente todas la ciencias de la vida y promete ser una solución a muchos problemas de relativa cotidianidad, la velocidad de producción de datos aventaja cuantitativamente al poder de análisis e interpretación que poseemos. Incluso nos vemos amenazados ante la emergencia de nuevas tecnologías cuando apenas nos hemos adaptado a las que recientemente se han transformado en estándares. Por estas razones, no siento más que satisfacción por haber contribuído mediante esta Tesis al desarrollo de esta disciplina emergente que, sin dudas, aún se encuentra en pleno crecimiento y depara grandes desafíos en el futuro cercano.

- FIN -