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

PhD Thesis

Comparative Genomics and Pan-Genomic Study of genus *Corynebacterium*

PH.D. STUDENT: **Luis Carlos Guimarães**

SUPERVISOR: Prof. Dr. Vasco Ariston de Carvalho Azevedo

CO-SUPERVISOR: Dr. Siomar de Castro Soares

BELO HORIZONTE

February – 2015

Luis Carlos Guimarães

Comparative Genomics and Pan-Genomic Study of genus *Corynebacterium*

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

SUPERVISOR: Prof. Dr. Vasco Ariston de Carvalho Azevedo

CO-SUPERVISOR: Dr. Siomar de Castro Soares

BELO HORIZONTE

February – 2015

043 Guimarães, Luis Carlos.
Comparative genomics and pan-genomic study of genus Corynebacterium [manuscrito] / Luis Carlos Guimarães. – 2015.

118 f. : il. ; 29,5 cm.

Orientador: Vasco Ariston de Carvalho Azevedo. Co-orientador: Siomar de Castro Soares.

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

1. Corynebacterium - Teses. 2. Genômica comparativa – Teses. 3. Pan-genoma. 4. Genoma central. 5. Filogenômica. 6. Genética – Teses. I. Azevedo, Vasco Ariston de Carvalho. II. Soares, Siomar de Castro. III. Universidade Federal de Minas Gerais. Departamento de Biologia Geral. IV. Título.

CDU: 575

Pós-Graduação em Genética
Departamento de Biologia Geral, ICB
Universidade Federal de Minas Gerais
Av. Antônio Carlos, 6627 - C.P. 486 - Pampulha - 31270-901 - Belo Horizonte - MG
e-mail: pg-gen@icb.ufmg.br FAX: (+31) - 3409-2570

Pós-Graduação

"Comparative genomics and pan-genomic study of genus corynebacterium"

Luis Carlos Guimarães

Tese aprovada pela banca examinadora constituída pelos Professores:

Dr. Vasco Ariston de Carvalho Azevedo - Orientador
UFMG

Dr. Siomar de Castro Soares - Coorientador
UFMG

Dr. Rommel Thiago Jucá Ramos
UFPA

Dr. Arthur Gruber
USP

Gabriel da Rocha Fernandes

Gabriel da Rocha Fernandes
FIOCRUZ

José Miguel Ortega
UFMG

Belo Horizonte, 11 de fevereiro de 2015.

I dedicate this work to my family, friends and every single person who believed and supported me, during my PhD.

ACKNOWLEDGEMENTS

“The all thoughts construction and scientific attitude takes years and is surrounded by several people that present their experiences, ideas barriers and solutions.”

With this manuscript I arrive to the end of another cycle in my life, for the beginning of another one.

It is for consideration and respect that I will not say thank nominally to each person who contributed to the development of this doctoral thesis. Even because, some of these people are not aware of my existence, the most people which I know do not notion of the value that represented in this process (even to push me up or push me down), those who know they have done a significant role in my life, probably never will read this manuscript and if happen to have access to it, can feel overvalued by the acknowledgements expressions.

This way, I must acknowledge and be thankful to all those people who in one way or another way were my "SUPERVISORS". I thank those who have guided my decision to be a biologist and after this, geneticist and bioinformatician. To my counselors: family, personal, emotional, professional, and intellectual, and why not cite, who guided me philosophically and spiritually, after all, as I once read in a text: "Does not matter if a researcher has or not religion, which he cannot deny is that research is a process of faith."

To all people that have guided my life journey to this point, I would like to say thank you so much and, always keep in mind that I will continue committed to honor my word to move forward in faith that I have in my work to transform it in exactly what it intended, a public utility product.

Special acknowledgements

Even no saying names before, I could not leave to say thank nominally to:

- Prof. Dr. Vasco Ariston de Carvalho Azevedo, supervisor of this PhD thesis, for all commitment, time expended during my own development and development of this work. As well as, non-academic advices and, also, for the friendship.

- Dr. Siomar de Castro Soares, co-supervisor of this PhD thesis, for all patience and always willing to help me; for making more pleasurable this period with his jokes and mainly for the friendship.

“All research is a permanent start and restart in converging cycles that represent the personal expression increasingly free, productive and constructive towards the benefit of all.”

Cerato SMM.

Table of contents

List of Figures

List of Tables

Abbreviations

AQUACEN	Laboratório Oficial Central do Ministério da Pesca e Aquicultura Ministério da Pesca e Aquicultura (National Reference Laboratory for Aquatic Animal Diseases)
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Coordination for the Improvement of Higher Education Personnel)
CeBiTec	Center for Biotechnology
CLIB	Cluster Industrial Biotechnology
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico (National Counsel of Technological and Scientific Development)
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DT	Diphtheria Toxin
EDTA	Ethylenediamine tetraacetic acid
Fapemig	Fundação de Amparo à Pesquisa do Estado de Minas Gerais (Foundation for Research Support of the State of Minas Gerais)
LGCM	Laboratório de Genética Celular e Molecular (Laboratory of Molecular and Cellular Genetics)
LPDNA	Laboratório do Polimorfismo do DNA (Laboratory of DNA Polymorphisms)
NGS	Next-Generation Sequencing
PLD	Phopholipase D
UFMG	Universidade Federal de Minas Gerais (Federal University of Minas Gerais)

Abstract

The genomic area has distinguished itself between different areas of modern biology, mainly after the advent of new high-throughput sequencing technologies (Next-Generation Sequencing - NGS), which allowed the massive deposit of whole genomes in public databases. The deluge of sequenced data, then, required the development of new bioinformatics software, with the ability to analyze these data, culminating in the creation of new areas such as comparative genomics that involves the comparison of the genetic content of an organism against another and helping the prediction of gene function and coding region sequences, identification of evolutionary events and finding of phylogenetic relationships. In this scenario, pan-genome studies allow the comparison of a large number of related bacteria enabling infer correlated genes with a lifestyle, gene repertoires and minimal genome size. In this work, first we describe the genome sequence, assembly and annotation of the *Corynebacterium urealyticum* strain DSM7111 followed by the comparison of two *C. urealyticum*, an opportunistic pathogen normally isolated from skin and mucous membranes in humans. Altogether, the strains encode 2,115 non-redundant coding sequences, of which 1,823 are commonly shared by both. Additionally, we applied the Vaxign software and we found 19 putative antigenic proteins through the application of the reverse vaccinology approach. Moreover, we expanded the comparative genomic analyses to the genus *Corynebacterium*, using 44 genomes of 25 different species. The pan-genomic analyses revealed an “open” pan-genome, which is currently composed of 22,177 protein-coding genes; the predicted core genome is composed of 562 genes; and, the predicted singletons account for 8,762 genes. Phylogenomic analysis using different strategies revealed the occurrence of four different clusters in the taxonomic tree: one composed of pathogenic, one of non-pathogenic and two of opportunistic pathogenic species. Finally, we presented genomes of *C. ulcerans*, which will be importante for future pan-genomic studies of this species.

I. Presentation

I.1 Collaborators

This work was executed at Laboratory of Molecular and Cellular Genetics (LGCM) – Federal University of Minas Gerais (UFMG); National Reference Laboratory for Aquatic Animal Diseases (AQUACEN) – Federal University of Minas Gerais (UFMG); Laboratory of DNA Polymorphisms (LPDNA) – Federal University of Pará; and Center for Biotechnology (CeBiTec) – Bielefeld University. In collaboration between the following researchers:

- Prof. Dr. Vasco Ariston de Carvalho Azevedo, Researcher and Professor from LGCM/UFMG, Brazil;
- Prof. Dr. Henrique César Pereira Figueiredo, Researcher and Professor from AQUACEN/UFMG, Brazil;
- Dr. Siomar de Castro Soares, Researcher from AQUACEN/UFMG, Brazil;
- Prof. Dr. Artur Luiz da Costa da Silva, Researcher and Professor from LPDNA/UFPA, Brazil;
- Prof. Dr. Rommel Thiago Jucá Ramos, Researcher and Professor from LPDNA/UFPA, Brazil;
- PD. Dr. Andreas Tauch, Researcher from CeBiTec and member of the Graduate Cluster Industrial Biotechnology (CLIB), Germany.

The work was supported by: Fundação de Amparo à Pesquisa do Estado de Minas Gerais (Fapemig), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and Cluster Industrial Biotechnology (CLIB).

II. Preface

II.1 Manuscript Structure

The thesis is divided in five chapters based on 1 review article (introduction) and 7 research articles, as follow:

a. The first chapter (introduction) presents the review article showing an overview about Pan-genomic studies. In this article, we start writting about the advent of new sequencing technologies – Next-Generation Sequencing (NGS), which enabled the massive submission of genomic data in public database. Consequently, the development of new bioinformatics areas as comparative genomics and subsequently pan-genomic studies – Submitted article to Current Genomics in 27-Jan-2015.
b. The second chapter presents a research article announcing the sequencing, assembling, annotation and deposit at GenBank of the *Corynebacterium urealyticum* DSM 7111 genome, isolated from urine samples of a 9-year-old patient with an ectopic kidney. The genome project is complete; under accession number NC_020230.1. The assembly, annotation and deposit of the genome data were done during my sandwich PhD – Published article in Genome Announcements.
c. The third chapter presents a research article comparing two genome sequences of *C. urealyticm* strains DSM 7109 and DSM 7111 cited above. The *C. urealyticum* DSM 7109 was sequenced, assembled and deposited by Andreas Tauch and colleagues under accession number NC_010545.1 at GenBank. In this work, we used comparative genomics strategies, analysis of metabolic pathways and, genome plasticity as well as, prediction of putative antigenic targets – Article accepted for publication in BMC Genomics.
d. The fourth chapter presents a research article describing the pan-genomic study of the genus *Corynebacterium*. In this work, we used 44 genomes of 25 different species with medical, veterinary and biotechnological importance. The pan-genome, core genome and singletons analyses were performed. All data were retrieved from Genbank and incorporated in EDGAR software. The statistical analyses were performed by R package – Submitted article to BMC Genomics in 27-oct-2014.
e. The fifth chapter is divided in six sections: the first section consists in a brief introduction about the *Corynebacterium ulcerans* and the state of the art; the second section describe the available data; and from the third to sixth sections we

present four research articles announcing the sequencing, assembly, annotation and deposit at GenBank of the four *Corynebacterium ulcerans,* strains 210932 (Published article in Genome Announcements), 210931 (Submitted article to SIGS in 15-oct-2014), FRC11 (Submitted article to Genome Announcements in 30-jan-2015) and 05146 (Submitted article to SIGS in 02-feb-2015), all isolated from humans. The genome projects are complete; under accession number CP009500.1, CP009583.1, CP009622.1 and CP009716.1, respectively. These works are the basis for the *C. ulcerans* pan-genome.

After the chapters, we present the general discussion and conclusions of the work. Finally, after bibliography, there is an "appendices" section, where one can find the curriculum vitae.

III. Introduction

III.1. Chapter I – Inside the pan-genome - methods and software overview[1]

Luis Carlos Guimarães, Leandro Benevides de Jesus, Marcus Vinícius Canário Viana, Rommel Thiago Jucá Ramos, Artur Silva, Siomar de Castro Soares, Vasco Azevedo.

Considering the importance of pan-genome studies, which allow a better understanding of the bacteria lifestyle through comparative genomics, and in view of the lack of a condensed material to explain this, we wrote an overview about this topic. The overview gives us a brief introduction on the comparative genomics emergence allowed by the Next-Generation Sequencing – NGS advent. This new sequencing platform boosted the development of comparative genomics and, consequently, the rising of the pan-genomic area through genomic sequencing of a large number of genomes from different isolates of the same organism allowing researchers to investigate several genomic features intrinsic to a given species. Additionally, we reviewed pan-genome concepts as core genome, accessory or dispensable genome and, species-specific or strain-specific genes, as well as, “open” and “closed” pan-genome. Furthermore, we summarize pan-genome results of free-living, facultative intracellular and obligate intracellular bacteria. Finally, we described software developed to calculate these analyses.

[1] Article accepted for publication in Current Genomics.

IV. Research Articles

IV.1. Chapter II. Complete genome sequence of Corynebacterium urealyticum strain DSM 7111, isolated from a 9-year-old patient with alkaline encrusted cystitis[2]

Luis Carlos Guimarães, Siomar Castro Soares, Andreas Albersmeier, Jochen Blom, Sebastian Jaenicke, Vasco Azevedo, Francisco Soriano, Andreas Tauch, Eva Trost

Corynebacterium urealyticum is an opportunistic pathogen causing mainly acute or encrusted cystitis, encrusted pyelitis, and pyelonephritis. This bacterium is normally isolated from the skin of hospitalized patients who are receiving broad-spectrum antibiotics. The sequencing of this strain will give us more data to perform pan-genomic study and try to understand the different lifestyle present in genus *Corynebacterium.* In this section we present an article announcing the sequencing, assembly, annotation and deposit at GenBank of the *Corynebacterium urealyticum* DSM 7111 genome, isolated from urine samples of a 9-year-old patient with an ectopic kidney. The genome project is complete; under accession number NC_020230.1.

[2] Published article in Genome Announcements.

IV.2. Chapter III Genome informatics and vaccine targets in *Corynebacterium urealyticum* using two whole genomes, comparative genomics, and reverse vaccinology[3]

Luis Carlos Guimarães, Siomar de Castro Soares, Eva Trost, Jochen Blom, Rommel Thiago Jucá Ramos, Artur Silva, Debmalya Barh, Vasco Azevedo

As highlighted in previous section about the importance of the comparative genomics studies in this article we present two genome comparison of multi-drug resistant *Corynebacterium urealyticum* (strains DSM 7109 and DSM 7111). Even both strains have been isolated from patients with alkaline encrusted cystitis, the strain DSM 7109 shows approximately 50 Kb more than strain DSM 7111 and 2,011 and 1,927 protein coding regions respectively. Together, both strains encode 2,115 non-redundant coding sequences, share in common 1,823 protein coding regions, the strain DSM 7109 has 188 strain-specific genes and the strain DSM7111 has 104 strain-specific genes. Additionally, reverse vaccinology analysis showed 19 putative antigenic proteins, among which the *spaDEF* operon that encodes pili forming proteins which may play a pivotal role in facilitating the adhesion of the pathogen to the host tissue.

[3] Article accepted for publication in BMC Genomics.

IV.3. Chapter IV A pan-genomic view of the genus *Corynebacterium*[4]

Luis Carlos Guimarães, Christina Bomholt, Jochen Blom, Siomar Castro Soares, Mariana Passos Santana, Rommel Thiago Jucá Ramos, Pablo Henrique Caracciolo Gomes de Sá, Ulisses de Pádua Pereira, Maria Paula Cruz Schneider, Artur Silva, Anderson Miyoshi, Vasco Azevedo, Andreas Tauch

The following research article shows the first work using 25 different species of the *Corynebacterium* genus level, for better understanding and evaluating the evolutionary relationship between pathogenic, non-pathogenic, and opportunistic pathogenic species. Using EDGAR software we predicted an “open” pan-genome for the genus, where new genes will be added for each newly sequenced genome. Additionally, core genome and singleton analyses revealed a low number of core genes and a high number of singletons, which mainly comes from the fact that our study considered bacteria with different lifestyles. Tests using different strategies for phylogenomic analyses were done combining whole genome-based, single-gene, and gene concatenation. Altogether, we found four clusters based on their lifestyle: one pathogenic cluster, one non-pathogenic cluster, and two opportunistic pathogenic clusters.

[4] Submitted article to BMC Genomics in 27-oct-2014.

A pan-genomic view of the genus *Corynebacterium*

Luis Carlos Guimarães[1,2], Christina Bomholt[2], Jochen Blom[3], Siomar Castro Soares[4], Mariana Passos Santana[1], Rommel Thiago Jucá Ramos[5], Pablo Henrique Caracciolo Gomes de Sá[5], Ulisses de Pádua Pereira[6], Maria Paula Cruz Schneider[5], Artur Silva[5], Anderson Miyoshi[1], Vasco Azevedo[1,*§], Andreas Tauch[2§]

[1]Department of General Biology, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil
[2]Center for Biotechnology, Bielefeld University, Bielefeld, Germany
[3]Bioinformatics and Systems Biology, Justus-Liebig-University Giessen, Giessen, Germany
[4]Department of Preventive Veterinary Medicine, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil
[5]Department of Genetics, Federal University of Pará, Belém, Pará, Brazil
[6]Department of Genetics, Federal University of Uberlândia, Uberlândia, Brazil

* Corresponding author

§ These authors share the senior authorship

Email addresses:

LCG: luisguimaraes.bio@gmail.com

CB: christina.bomholt@uni-bielefeld.de

JB: jochen.blom@computational.bio.uni-giessen.de

SCS: siomars@gmail.com

MPS: santana.maripassos@gmail.com

RTJR: rommelthiago@gmail.com

PHCGS: pablogomesdesa@gmail.com

UPP: upaduapereira@gmail.com

MPCS: paula@ufpa.br

AS: asilva@ufpa.br

AM: miyoshi@icb.ufmg.br

VA: vasco@icb.ufmg.br

AT: tauch@cebitec.uni-bielefeld.de

Abstract

Background

The genus *Corynebacterium* represents a taxon of diverse bacteria from human, animal and environmental sources. It is characterized by a specific organization of the bacterial cell wall and by its high G+C content of the DNA. We report here a comparative genomic study using 44 genome sequences from members of the genus *Corynebacterium*.

Results

This *in silico* analysis revealed that the pan-genome is currently composed of 22,177 protein-coding genes, including 8,762 strain-specific genes (singletons). The mean number of singletons estimated to be added by each additionally sequenced corynebacterial genome comprises 236 protein-coding genes. The number of genes present in the core genome will comprise about 580 protein-coding genes when adding further genome sequences to the current data set. Considering that the genus *Corynebacterium* is composed of species with very different lifestyles, the low number of genes assigned to the core genome and the large number of predicted singletons indicate a conservation of genes with basic cellular functions and the acquirement of new genes that may be advantageous for the adaptation to new environments. Phylogenomic analysis using different strategies showed the presence of four distinct clusters in the taxonomic tree of the genus *Corynebacterium* one pathogenic, one non-pathogenic and two opportunistic pathogenic.

Conclusions

This way, the comparative genomic study of the *Corynebacterium* genus shows a foundation for identification and analysis of new genes that may be involved in more diverse adaptation to different hosts and environments.

Key words: *Corynebacterium*, Pan-genome, Core genome, Singletons, Phylogenomi

Background

The taxonomic class *Actinobacteria* is the third largest phylogenetic branch among the bacteria [1]. This taxon is of substantial scientific interest as it includes a large number of pathogenic species and numerous bacteria of significant relevance for biotechnology [2]. The so-called CMNR group of this taxon comprises the genera *Corynebacterium*, *Mycobacterium*, *Nocardia*, and *Rhodococcus* that are extensively studied due to their medical, veterinary and biotechnological importance. These four genera are characterized by a specific organization of the cell wall including mycolic acids and by their high G+C content of the DNA [3]. The genus *Corynebacterium* was initially created to include the pathogenic species *Corynebacterium diphtheriae*, the causative agent of diphtheria [4], but numerous other species were included later into this taxon that nowadays contains a large collection of corynebacteria with various lifestyles [2]. Hitherto sequenced corynebacterial genomes consist of a circular chromosome with a size ranging from 2.3 Mbp to 3.4 Mbp and of 2,000 to 3,200 genes (Table 1). So far, the genomes of species from human clinical sources and from animals have been shown to be smaller than those found in other ecosystems, such as soil and dairy products (Table 1). Moreover, the order of orthologous genes is highly conserved throughout the species [2], although *Corynebacterium jeikeium*, *Corynebacterium resistens*, *Corynebacterium urealyticum*, and *Corynebacterium variabile* show inversions and evidence for their common evolutionary origin [5–8].

A previous evolutionary study using *16S* rRNA and *rpoB* gene sequence showed a close relationship between *C. jeikeium*, *C. urealyticum*, and *C. variabile* [9]. However, besides the accuracy of *rpoB* analyses, a recent study using whole genome comparisons has shown a divergent location for the low-pathogenic potential species *C. kroppenstedtii* in the phylogenetic tree, where it appears between clusters of pathogenic and non-pathogenic species [10].

Moreover, there is a lack of phylogenetic information using other conserved genes, and also of genomic comparisons at a pan-genomic level that could correlate the lifestyle of all species with their gene conservation and their position in the phylogenetic tree. Thus, studies employing comparative genomics approaches to evaluate the lifestyle of *Corynebacterium* genus and the phylogenetic position of the species are warranted.

Table 1: General information about the 44 genomes of members of the genus *Corynebacterium.*

Species	Genome size (bp)	G+C content (mol%)	No. of genes	No. of Singletons	Origin of sequenced strain	GenBank accession No.	Status of project
C. accolens ATCC 49725 (op)	2,413,333	59.7	2,390	75	Human; respiratory tract	NZ_ACGD00000000.1	draft
C. accolens ATCC 49726 (op)	2,368,466	59.6	2,417	116	Human; stool of infant	NZ_AEED00000000.1	draft
C. ammoniagenes DSM 20306 (op)	2,763,612	55.6	2,703	268	Stool of infant	NZ_ADNS00000000.1	draft
C. amycolatum SK46 (op)	2,513,912	58.6	2,241	440	Human; skin	NZ_ABZU00000000.1	draft
C. aurimucosum ATCC 700975 (op)	2,790,189	60.5	2,630	200	Human; urogenital tract	NC_012590.1	complete
C. bovis DSM 20582 (op)	2,522,962	75.5	2,392	594	Cow; milk	NZ_AENJ00000000.1	draft
C. casei UCMA 3821 (np)	3,112,736	55.3	2,925	339	Cheese; surface	NZ_CAFW00000000.1	draft
C. diphtheriae 31A (p)	2,535,346	53.6	2,458	65	Human; pharyngeal region	NC_016799.1	complete
C. diphtheriae C7 (beta) (p)	2,499,189	53.5	2,415	87	Derivative of the avirulent isolate C7	NC_016801.1	complete
C. diphtheriae NCTC13129 (p)	2,488,635	53.5	2,388	137	Human; pharyngeal membrane	NC_002935.2	complete
C. diphtheriae PW8 (p)	2,530,683	53.7	2,414	48	Human; pharyngeal region	NC_016789.1	complete
C. durum F0235 (op)	2,809,096	56.8	2,877	874	Human; dental plaque	NZ_AMEM00000000.1	draft
C. efficiens YS-314 (np)	3,147,090	62.9	3,064	473	Soil; Japan	NC_004369.1	complete
C. genitalium ATCC 33030 (op)	2,349,653	62.7	2,290	359	Human; Urogenital tract	NZ_ACLJ00000000.2	draft
C. glucuronolyticum ATCC 51866 (op)	2,845,674	59.0	2,790	240	Human; semen	NZ_ACHF00000000.1	draft
C. glucuronolyticum ATCC 51867 (op)	2,784,713	59.2	2,704	183	Human; urogenital tract	NZ_ABYP00000000.1	draft
C. glutamicum ATCC 13032 (np)	3,309,401	53.8	3,073	88	Soil	NC_003450.3	complete
C. glutamicum ATCC 13032 (np)	3,282,708	53.8	3,138	29	Soil	NC_006958.1	draft
C. glutamicum ATCC 14067 (np)	3,226,741	54.1	3,139	218	Soil	NZ_AGQQ00000000.1	draft
C. glutamicum R (np)	3,314,179	54.1	3,156	108	Soil	NC_009342.1	complete
C. glutamicum S9114 (np)	3,262,889	53.9	3,069	128	Soil	NZ_AFYA00000000.1	draft
C. jeikeium ATCC 43734 (op)	2,426,461	61.6	2,280	158	Human; blood	NZ_ACYW00000000.1	draft
C. jeikeium K411 (op)	2,462,499	61.4	2,181	27	Human; skin	NC_007164.1	complete
C. kroppenstedtii DSM 44385 (op)	2,446,804	57.5	2,083	473	Human; sputum	NC_012704.1	complete
C. lipophiloflavum DSM 44291	2,293,743	64.8	2,430	334	Human; urogenital tract	NZ_ACHJ00000000.1	draft
C. matruchotii ATCC 14266 (op)	2,855,988	57.1	2,680	148	Human; oral flora	NZ_ACSH00000000.2	draft
C. matruchotii ATCC 33806 (op)	2,967,145	57.0	3,197	603	Human; gingival crevices	NZ_ACEB00000000.1	draft
C. nuruki S6-4 (np)	3,106,595	69.5	2,856	400	Alcohol fermentation	NZ_AFIZ00000000.1	draft
C. pseudogenitalium ATCC 33035 (op)	2,600,726	59.5	2,560	168	Human; urogenital tract	NZ_ABYQ00000000.2	draft
C. pseudotuberculosis 1002 (p)	2,335,113	52.2	2,203	2	Goat; caseous granulomas	NC_017300.1	complete
C. pseudotuberculosis C231 (p)	2,328,208	52.2	2,204	4	Sheep; abscess	NC_017301.1	complete
C. pseudotuberculosis CIP52.97 (p)	2,320,595	52.1	2,194	44	Horse; ulcerative lymphangitis	NC_017307.1	complete
C. pseudotuberculosis FRC41 (p)	2,337,913	52.2	2,171	14	Human; inguinal lymph node	NC_014329.1	complete
C. pseudotuberculosis I19 (p)	2,337,730	52.2	2,213	1	Cow, mastitis	NC_017303.1	complete
C. pseudotuberculosis PAT10 (p)	2,335,323	52.2	2,200	2	Sheep; abscess	NC_017305.1	complete
C. resistens DSM 45100 (op)	2,601,311	57.1	2,230	197	Human; blood	NC_015673.1	complete
C. striatum ATCC 6940 (op)	2,724,288	59.4	2,730	288	Human; urogenital tract	NZ_ACGE00000000.1	draft
C. tuberculostearicum SK141 (op)	2,372,261	60.0	2,266	83	Human; skin	NZ_ACVP00000000.1	draft
C. ulcerans 0102 (p)	2,579,188	53.4	2,417	89	Human; pharyngeal pseudomembrane	NC_018101.1	complete
C. ulcerans 809 (p)	2,502,095	53.3	2,246	23	Human; brochoalveolar sample	NC_017317.1	complete
C. ulcerans BR-AD22 (p)	2,606,374	53.4	2,402	55	Dog; nasal sample	NC_015683.1	complete
C. urealyticum DSM 7109 (op)	2,369,219	64.2	2,082	75	Human, bladder stone	NC_010545.1	complete
C. urealyticum DSM 7111 (op)	2,316,065	64.2	2,007	25	Human; urine	NC_020230.1	complete
C. variabile DSM 44702 (np)	3,433,007	67.1	3,131	480	Cheese, suface	NC_015859.1	complete

To better understand the lifestyle of *Corynebacterium* species and to evaluate the evolutionary relationship between them, we performed a comparative analysis with 44 genomes of bacteria isolated from several niches. Comparative genomics analyses in related bacteria have shown an extensive genomic intra-species diversity and highlighted bacterial promiscuity [11].

In the subsequent sections, we describe the pan-genome, the core genome and the singletons for the genus *Corynebacterium*. The pan-genome concept was used for the first time by Tettelin and colleagues in 2005 [12]. The main goal of pan-genomics is the genomic comparison of different members of the same taxon to get insights into whole gene repertoire and the genomic diversity of a given taxonomic group [12, 13]. The pan-genome involves: (i) the core genome formed by a subset of genes shared by all genomes and normally involved in essential cellular processes; (ii) the accessory genome composed of a subset of genes present in two or more genomes, but not in all; and (iii) the group of singletons comprising genes present in a single genome [12, 14]. Genes present in the accessory genome and singletons are usually involved in functions related to niche adaptation [13, 15].

At the genus level, this is the first pan-genomic work with 25 different corynebacterial species.

Results

Corynebacterium pan-genome subsets: pan-genome, core genome, and singletons

To calculate the total number of non-redundant genes (pan-genome) of the genus *Corynebacterium* (44 species), we used the software EDGAR. Briefly, this software analyzes the orthologous genes based on the SRV method [16] by using an iterative pairwise comparison. *C. glutamicum* R genome was selected as reference and its gene content was taken as the base set for comparisons with the total gene content of all the other species. The predicted *Corynebacterium* genus pan-genome subset is composed of 22,177 genes (Fig. 1), which is 8.64-fold the mean number of genes in each genome (2,537). The predicted size of the core genome subset is 562 genes, representing only 2.53% of the entire pan-genome. The singleton subset, which is composed of genes present in only one genome of the data set, therefore directly affecting the total number of genes in the pan-genome, is composed of 8,762 genes.

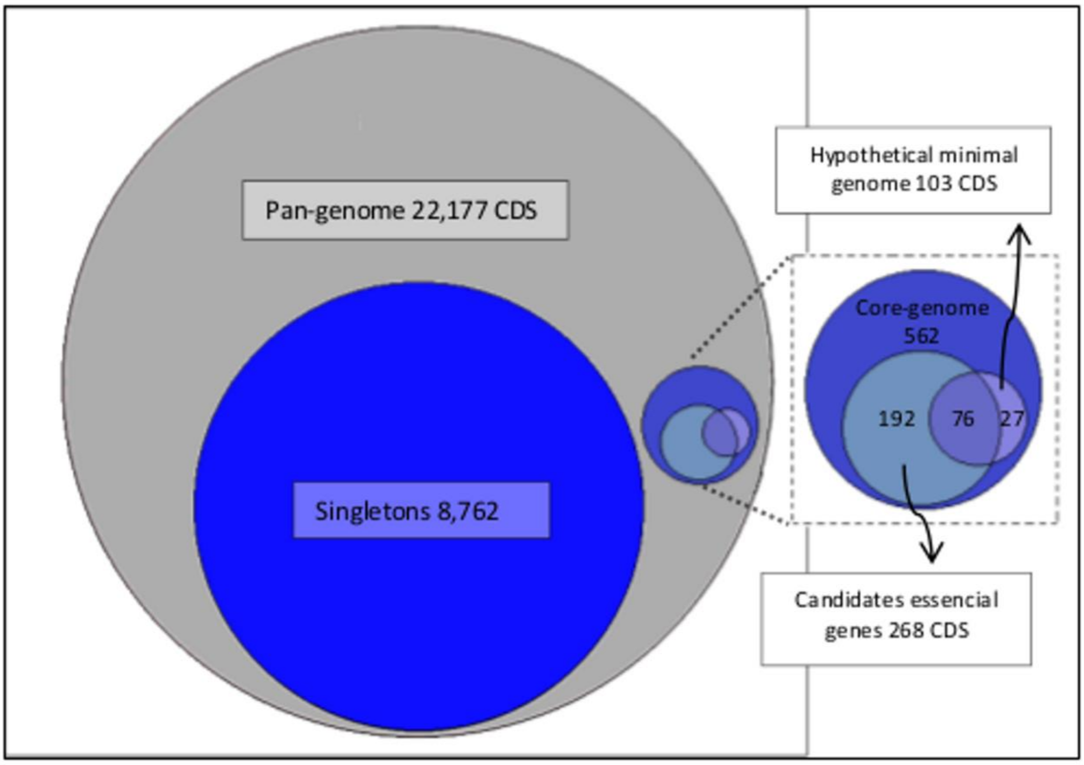
Pan-genome 22,177 CDS
Singletons 8,762
Hypothetical minimal genome 103 CDS
Core-genome 562
192
76
27
Candidates essencial genes 268 CDS

Fig. 1: Venn diagram representing pan-genome, core-genome, and singletons of the 44 genomes of *Corynebacterium* genus. Pan-genome, number of non-redundant genes composing *Corynebacterium* genus. Singletons, number of genes present in only one genome. Core genome, number of genes present in two or more genomes, but not in all. Hypothetical minimal genome and candidates for essential genes, number of genes of the core genome calculated using previously studies [17, 18].

Next, we compared the core-genome subset with 658 genes previously defined as candidate essential genes from *C. glutamicum* R [17] and 123 Coding DNA sequences (CDSs) from the core genome of Actinobacteria [2]. From those, 268 essential genes of *C. glutamicum* R and 103 actinobacterial essential CDSs are part of the total core genome of the genus *Corynebacterium*. Interestingly, 76 CDSs are shared between both datasets (Fig. 1).

Corynebacterium pan-genome extrapolations

An extrapolation of the corynebacterial pan-genome was calculated with 10,000 randomly selected permutations out of the set of 44 genomes using Heap's Law ($n = \kappa \times N^{\gamma}$). In this study, the variables κ and γ were determined to be 2,335.399 and 0.595, respectively, with N being the variable number of genomes (Fig. 2). According to Heap's Law, the pan-genome is considered closed when $\alpha > 1$ (using the formula $\alpha = 1 - \gamma$) and the addition of new genomes will not increase the total number of genes significantly. On the other hand, the pan-genome is open when $\alpha < 1$ and for each newly added genome the number of genes in the pan-genome will increase [13]. As we calculated $\alpha = 0.405$, it is possible to infer that the pan-genome for the 44 genomes of the genus *Corynebacterium* considered here is open.

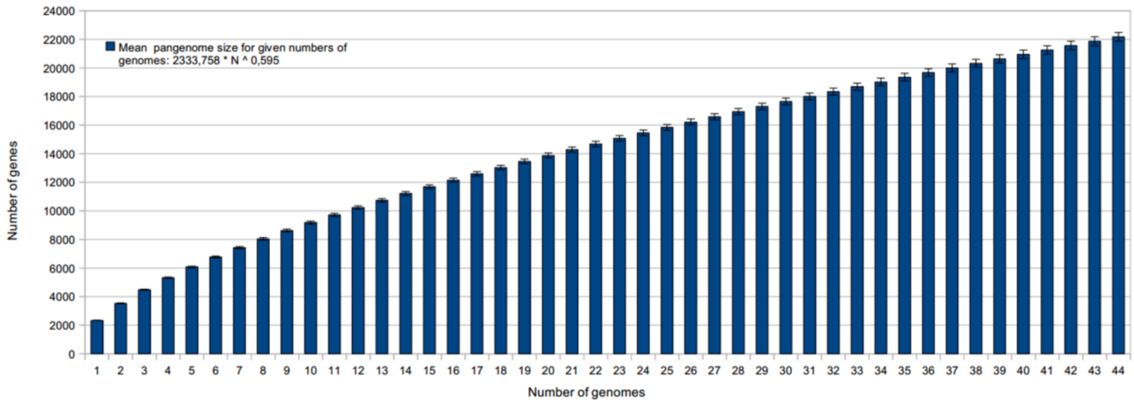

Fig. 2: Pan-genome development of the genus *Corynebacterium*. The pan-genome extrapolation is based on 10,000 randomly selected permutations of the set of 44 genomes belonging to members of the genus *Corynebacterium*.

The *Corynebacterium* core genome extrapolations

The extrapolation of the core genome subset was calculated using the formula obtained from the least-squares fit of the exponential regression decay to the mean values (Fig. 3). As a result, a *tg(θ)* of 580.921 was predicted, meaning that the corynebacterial core genome tends to stabilize at ~580 genes. Compared to the extrapolation of the pan-genome (Fig. 2), the curve of the core genome reached a stable plateau, indicating that the addition of new corynebacterial genomes will not significantly decrease the number of genes in the core genome (Fig. 3). Therefore, the current set of core genes is already representative of the final stabilizing core genome and provides a solid basis for phylogenomic studies of the genus *Corynebacterium*.

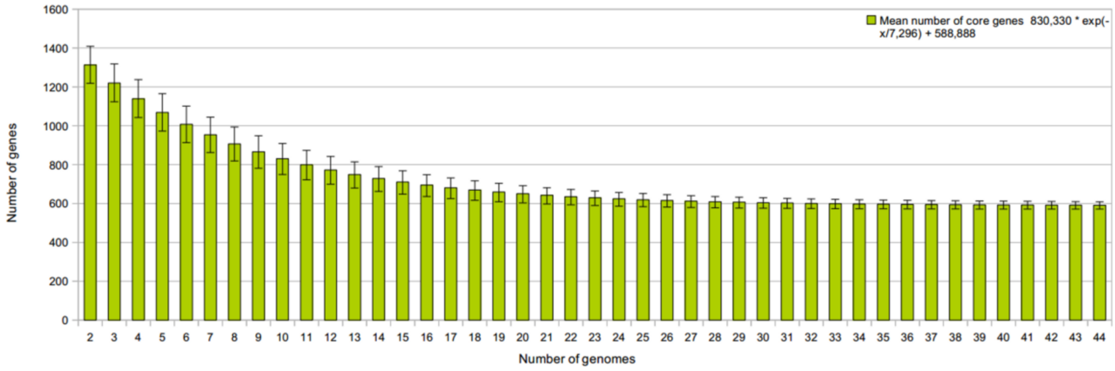

Fig. 3: Core genome development of the Corynebacterium genus. The core genome extrapolation is based on 10,000 randomly selected permutations of the set of 44 genomes belonging to members of the genus Corynebacterium.

Corynebacterium singleton extrapolations

We also performed extrapolation analysis for singletons, as described in the Materials and Methods section. The *tg(θ)* was predicted to be 235.619, meaning that, for each new genome, ~236 genes will be added to the total pan-genome (Fig. 4). Again, the extrapolation curve reaches a stability plateau, suggesting that the addition of new genomes will not significantly change the number of singleton genes.

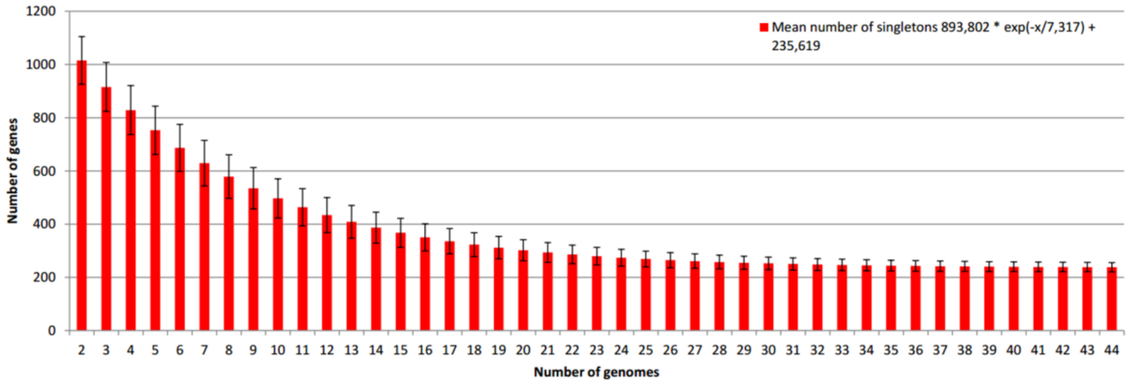

Fig. 4: Singleton development of *Corynebacterium* genus. The singleton extrapolation is based on 10,000 randomly selected permutations of the set of 44 genomes belonging to members of the genus Corynebacterium.

Pathogenic, non-pathogenic, and opportunistic *Corynebacteria* core genomes

The 44 species studied herein were divided into pathogens, opportunistic pathogens, and non-pathogens according to the literature (Table 1); and the differential core genome, comprising the specific core genes of a given subset that are not present in the core genome of the 44 species, was calculated for each one using the SRV method with EDGAR software. The differential core genome of pathogenic *Corynebacteria* contains 272 genes using *C. pseudotuberculosis* FRC41 as reference; the core genome from opportunistic *Corynebacteria* contains 125 genes using *C. urealyticum* 7109 as reference; and the subset from non-pathogens contains 462 genes using *C. glutamicum* R as reference.

The core genome of the total dataset and the differential core genome of the pathogens, opportunistic pathogens, and non-pathogens were analyzed by COG functional category. According to figure 5, a large number of genes for the core genome were related to the “Metabolism” category, followed by the “Information Storage and Processing” category. Only a few genes are included in the “Poorly Characterized” category. For the differential core genomes, a high number of genes were related to the “Metabolism” and “Poorly Characterized” categories and a low number to “Information Storage and Processing.”

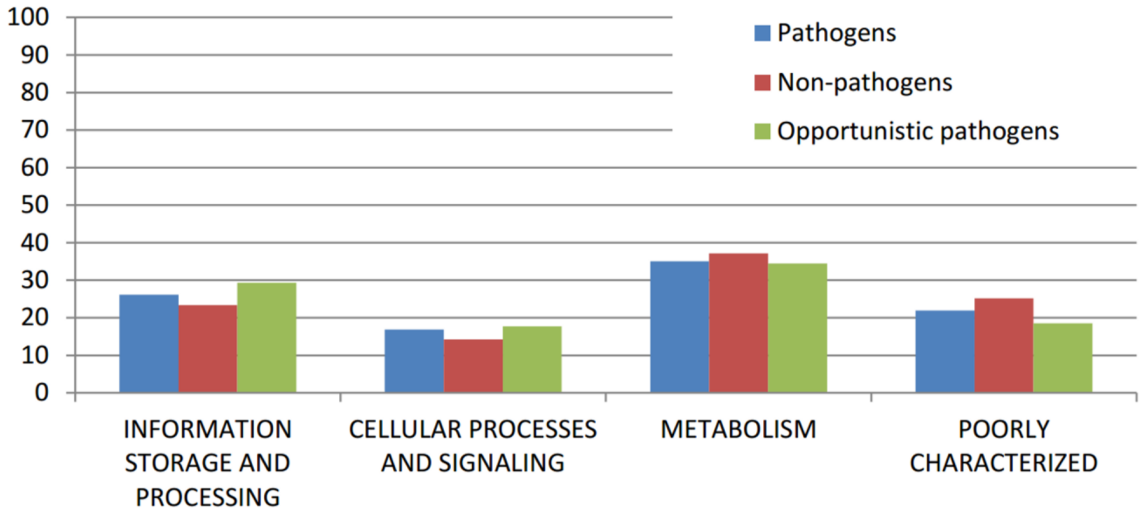

Fig. 5: Core genes of the *Corynebacterium* genus classified by COG functional category. Core genome all species, the genes composing the core genome of 44 genomes; Differential core genome pathogens, the genes of the core genome of pathogen species which were absent in one or more genomes of the non-pathogens and opportunistic pathogens; Differential core genome opportunistic pathogens, the genes of the core genome of opportunistic pathogen species which were absent in one or more genomes of the non-pathogens and pathogens; Differential core genome non-pathogens, the genes of the core genome of non-pathogen species which were absent in one or more genomes of the pathogens and opportunistic pathogens.

Corynebacterium phylogenomics

Phylogenetic analyses were performed by using single-gene (*16S* rRNA and *rpoB*), gene concatenation (two porin genes and five conserved proteins), and whole-genome datasets to compare the resolution of the five resulting phylogenetic trees. We used the Maximum Likelihood method with MEGA software [19] to construct the single-gene and gene concatenation trees. For the whole-genome, we used the EDGAR’s pipeline [16].

The shared gene content from the 44 genomes was predicted with the EDGAR software and each genome was cross-compared to plot a heatmap using the Gegenees software (Fig. 6). The pathogenic species *C. diphtheriae*, *C. ulcerans*, and *C. pseudotuberculosis* formed a related cluster. The pathogenic species *C. kroppenstedtii* and *C. bovis* were positioned in a cluster closely related to two non-pathogenic species (*C. nuruki* and *C. variabile*) and four opportunistic species (*C. amycolatum*, *C. jeikeium*, *C. urealyticum*, and *C. resistens*). *C. glutamicum* and *C. efficiens*, also non-pathogenic species of great industrial importance as amino acid producers, were closely related in a different cluster. The opportunistic pathogenic species *C. matruchotii* and *C. durum* were positioned between pathogenic and non-pathogenic species. Finally, the opportunistic pathogenic species *C. jeikeium*, *C. resistens*, and *C. urealyticum* were also positioned between pathogenic and non-pathogenic species, although in a different cluster. All other opportunistic pathogenic species formed different clusters.

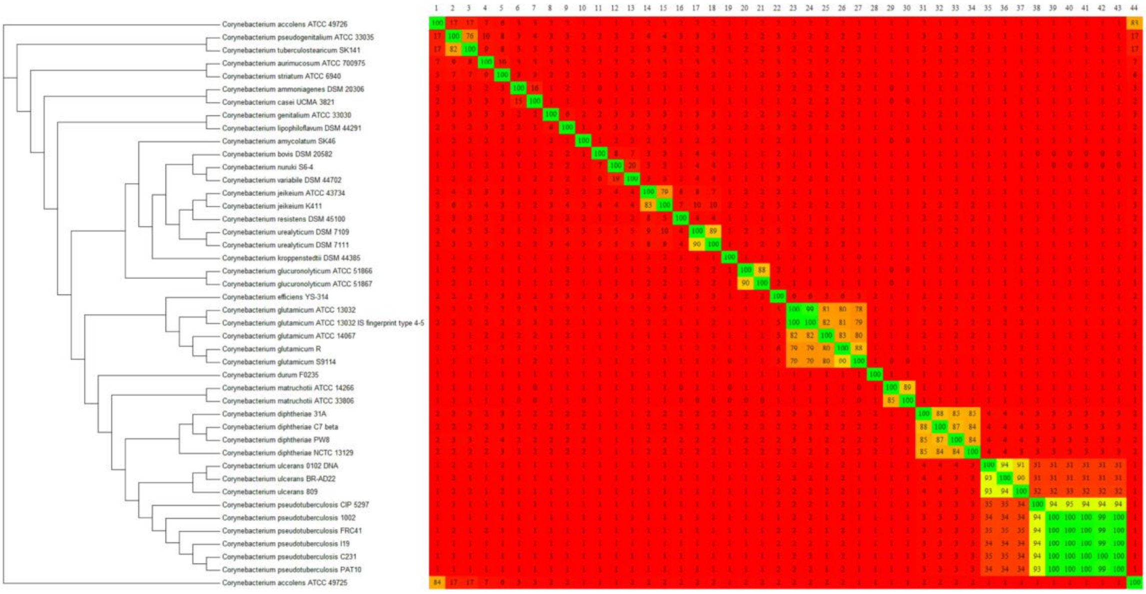

Fig. 6: Phylogenomic tree and heatmap analyses of the *Corynebacterium* genus. Comparisons between the genomes were plotted as percentages of similarity on the heatmap using Gegenees (version 1.1.4). The percentage of similarity was used to generate a phylogenomic tree with MEGA (version 5.10). Numbers from 1 to 44 represent species from *C. accolens* ATCC 49726 to *C. accolens* ATCC 49725 (upper-left to lower-left corner). Percentages were plotted with a spectrum ranging from red (low similarity) to green (high similarity).

The tree based on *16S rRNA* gene (Fig. 7) also showed good resolution but, different from the tree based on the whole-genome, *C. diphtheriae* NCTC 13129 was positioned in a cluster together with opportunistic pathogenic species (*C. matruchotti*, *C. accolens*, *C. tuberculostearicum*, and *C. pseudogenitalium*) instead of clustering with other pathogenic species (*C. diphtheriae*, *C. ulcerans*, *C. pseudotuberculosis*). In agreement with the whole-genome tree, the pathogenic species *C. kroppenstedtii* and *C. bovis* were also positioned among opportunistic (*C. amycolatum*, *C. jeikeium*, *C. urealyticum*, and *C. resistens*) and non-pathogenic species (*C. variabile* and *C. nuruki*). The non-pathogenic species *C. casei* and *C. glutamicum S9114* were positioned in a distant unrelated cluster among opportunistic pathogenic species. The *C. durum* opportunistic pathogenic species was positioned among non-pathogenic species.

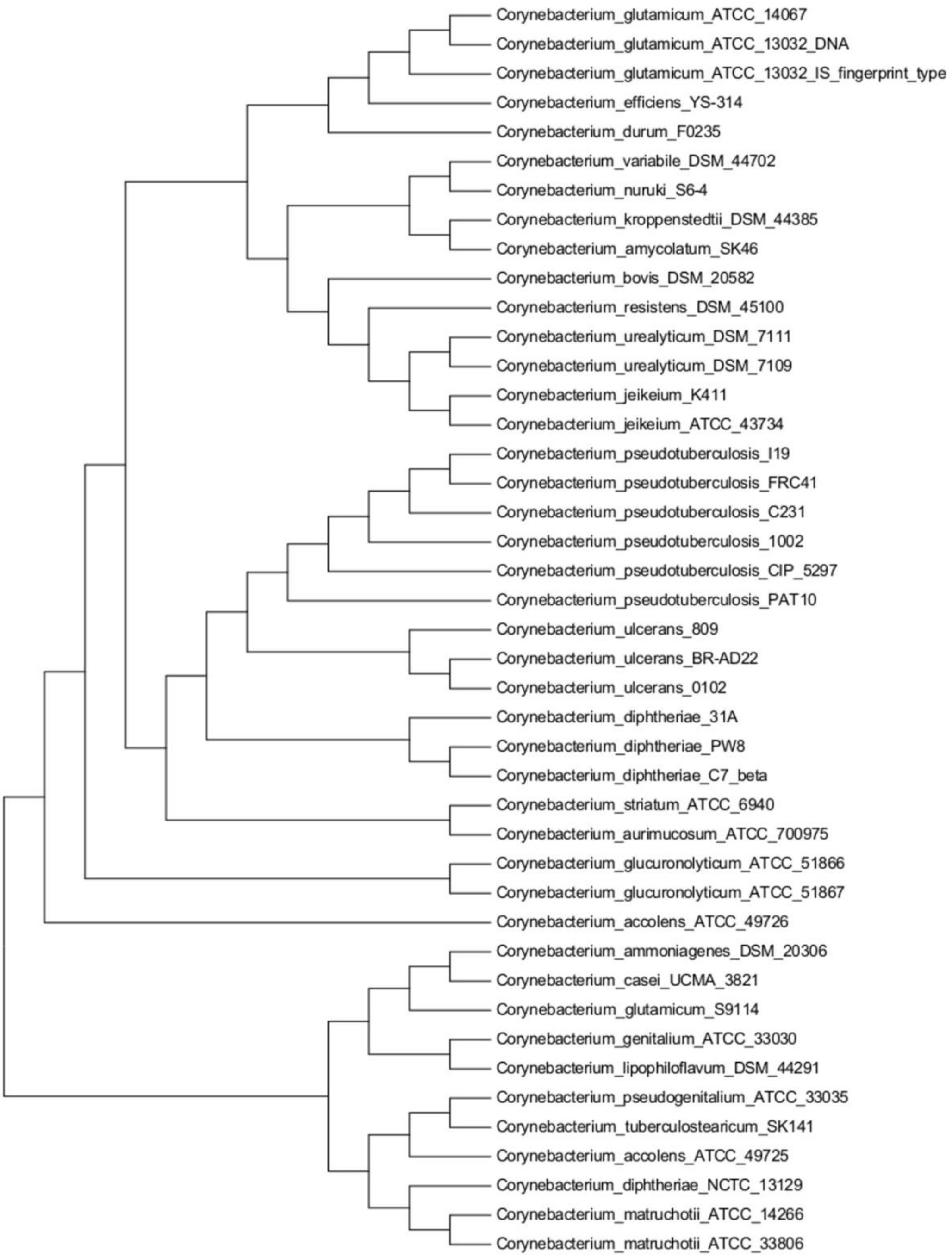
Corynebacterium_glutamicum_ATCC_14067
Corynebacterium_glutamicum_ATCC_13032_DNA
Corynebacterium_glutamicum_ATCC_13032_IS_fingerprint_type
Corynebacterium_efficiens_YS-314
Corynebacterium_durum_F0235
Corynebacterium_variabile_DSM_44702
Corynebacterium_nuruki_S6-4
Corynebacterium_kroppenstedtii_DSM_44385
Corynebacterium_amycolatum_SK46
Corynebacterium_bovis_DSM_20582
Corynebacterium_resistens_DSM_45100
Corynebacterium_urealyticum_DSM_7111
Corynebacterium_urealyticum_DSM_7109
Corynebacterium_jeikeium_K411
Corynebacterium_jeikeium_ATCC_43734
Corynebacterium_pseudotuberculosis_I19
Corynebacterium_pseudotuberculosis_FRC41
Corynebacterium_pseudotuberculosis_C231
Corynebacterium_pseudotuberculosis_1002
Corynebacterium_pseudotuberculosis_CIP_5297
Corynebacterium_pseudotuberculosis_PAT10
Corynebacterium_ulcerans_809
Corynebacterium_ulcerans_BR-AD22
Corynebacterium_ulcerans_0102
Corynebacterium_diphtheriae_31A
Corynebacterium_diphtheriae_PW8
Corynebacterium_diphtheriae_C7_beta
Corynebacterium_striatum_ATCC_6940
Corynebacterium_aurimucosum_ATCC_700975
Corynebacterium_glucuronolyticum_ATCC_51866
Corynebacterium_glucuronolyticum_ATCC_51867
Corynebacterium_accolens_ATCC_49726
Corynebacterium_ammoniagenes_DSM_20306
Corynebacterium_casei_UCMA_3821
Corynebacterium_glutamicum_S9114
Corynebacterium_genitalium_ATCC_33030
Corynebacterium_lipophiloflavum_DSM_44291
Corynebacterium_pseudogenitalium_ATCC_33035
Corynebacterium_tuberculostearicum_SK141
Corynebacterium_accolens_ATCC_49725
Corynebacterium_diphtheriae_NCTC_13129
Corynebacterium_matruchotii_ATCC_14266
Corynebacterium_matruchotii_ATCC_33806

Fig. 7: Phylogenetic tree based on *16S* rRNA. The construction was obtained by the Maximum Likelihood statistical method. The tree was derived from the alignments of *16S* rRNA gene sequences. The phylogenetic distances were calculated by the software MEGA 5.

The phylogenetic tree based on *rpoB* (RNA polymerase β-subunit) gene (Fig. 8) shows very good resolution and agreement with the whole-genome phylogenetic tree. Indeed, the non-pathogenic species formed a closely related cluster, except for the non-pathogenic species *C. casei*. Moreover, in both cases (*rpoB* and Gegenees), *C. casei* was positioned in a different cluster among opportunistic pathogenic species (*C. ammoniagenes*, *C. lipophiloflavum*, *C.*

genitalium, *C. aurimucosum*, *C. striatum*, *C. accolens*, *C. pseudogenitalium*, *C. tuberculosteriacum*); and *C. nuruki* and *C. variabile* formed another cluster with pathogenic (*C. bovis*) and opportunistic pathogenic (*C. urealyticum, C. resistens*, and *C. jeikeium*) species. The pathogenic species *C. pseudotuberculosis, C. ulcerans*, and *C. diphtheriae* formed a closely related cluster.

The proteins *Isoleucyl tRNA synthetase*, *ribosomal protein S1*, *DNA topoisomerase*, *SecY*, and *GTPase* are highly conserved in *Corynebacterium* species [20]. The phylogenetic tree based on the concatenation of the predicted sequence of those proteins also produced a good resolution (Fig. 9) and agreement with the whole-genome plylogenetic tree. The pathogenic species *C. pseudotuberculosis*, *C. ulcerans*, and *C. diphtheriae* formed a closely related cluster. The pathogenic species *C. kroppenstedii* and *C. bovis* were positioned in a different cluster among opportunistic pathogenic species (*C. amycolatum*, *C. urealyticum, C. resistens*, and *C. jeikeium*) and non-pathogenic species (*C. nuruki* and *C. variabile*). The non-pathogenic species *C. glutamicum* and *C. efficiens* formed a closely related cluster, and *C. casei* (another non-pathogenic species) was positioned in a different cluster among opportunistic pathogenic species.

The phylogenetic tree based on the concatenation of two porins, *porA* and *porH*, gives a good resolution; however, the species arrangement is in disagreement with the above presented phylogenetic trees (Fig. 10). The pathogenic species *C. pseudotuberculosis* and *C. ulcerans* formed a cluster which is also closely related to the opportunistic pathogens *C. jeikeium* and *C. resistens*. Unexpectedly, *C. diphtheriae* strains split into one cluster with opportunistic pathogens (*C. striatum*, *C. aurimucosum*, and *C. pseudogenitalium*) and another with non-pathogenic (*C. casei*) and opportunistic pathogens (*C. durum*, *C. ammoniagenes*, and *C. matruchotii*). The non-pathogenic species *C. glutamicum* and *C. efficiens* formed a cluster closely related to an opportunistic pathogen (*C. urealyticum*) whereas *C. casei* (another non-pathogenic species) was positioned in a different cluster among pathogenic and opportunistic pathogenic species.

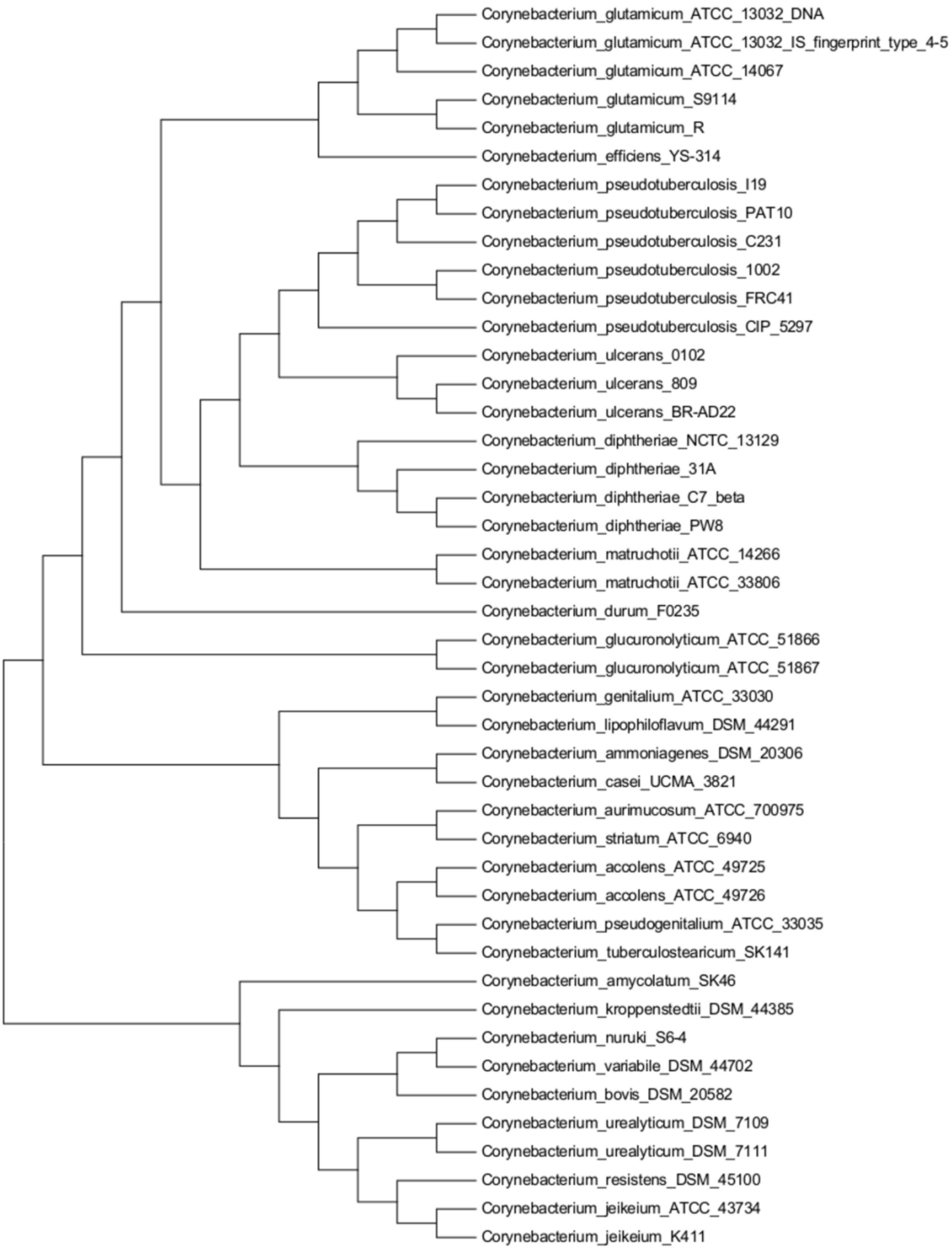
Corynebacterium_glutamicum_ATCC_13032_DNA
Corynebacterium_glutamicum_ATCC_13032_IS_fingerprint_type_4-5
Corynebacterium_glutamicum_ATCC_14067
Corynebacterium_glutamicum_S9114
Corynebacterium_glutamicum_R
Corynebacterium_efficiens_YS-314
Corynebacterium_pseudotuberculosis_I19
Corynebacterium_pseudotuberculosis_PAT10
Corynebacterium_pseudotuberculosis_C231
Corynebacterium_pseudotuberculosis_1002
Corynebacterium_pseudotuberculosis_FRC41
Corynebacterium_pseudotuberculosis_CIP_5297
Corynebacterium_ulcerans_0102
Corynebacterium_ulcerans_809
Corynebacterium_ulcerans_BR-AD22
Corynebacterium_diphtheriae_NCTC_13129
Corynebacterium_diphtheriae_31A
Corynebacterium_diphtheriae_C7_beta
Corynebacterium_diphtheriae_PW8
Corynebacterium_matruchotii_ATCC_14266
Corynebacterium_matruchotii_ATCC_33806
Corynebacterium_durum_F0235
Corynebacterium_glucuronolyticum_ATCC_51866
Corynebacterium_glucuronolyticum_ATCC_51867
Corynebacterium_genitalium_ATCC_33030
Corynebacterium_lipophiloflavum_DSM_44291
Corynebacterium_ammoniagenes_DSM_20306
Corynebacterium_casei_UCMA_3821
Corynebacterium_aurimucosum_ATCC_700975
Corynebacterium_striatum_ATCC_6940
Corynebacterium_accolens_ATCC_49725
Corynebacterium_accolens_ATCC_49726
Corynebacterium_pseudogenitalium_ATCC_33035
Corynebacterium_tuberculostearicum_SK141
Corynebacterium_amycolatum_SK46
Corynebacterium_kroppenstedtii_DSM_44385
Corynebacterium_nuruki_S6-4
Corynebacterium_variabile_DSM_44702
Corynebacterium_bovis_DSM_20582
Corynebacterium_urealyticum_DSM_7109
Corynebacterium_urealyticum_DSM_7111
Corynebacterium_resistens_DSM_45100
Corynebacterium_jeikeium_ATCC_43734
Corynebacterium_jeikeium_K411

Fig. 9: Phylogenetic tree based on 5 conserved proteins. The construction was obtained by the Maximum Likelihood statistical method. The tree was derived from the alignments of *isoleucyl tRNA synthetase*, *ribosomal protein S1*, *DNA topoisomerase*, *SecY*, and *GTPase* sequences. The phylogenetic distances were calculated by the software MEGA 5.

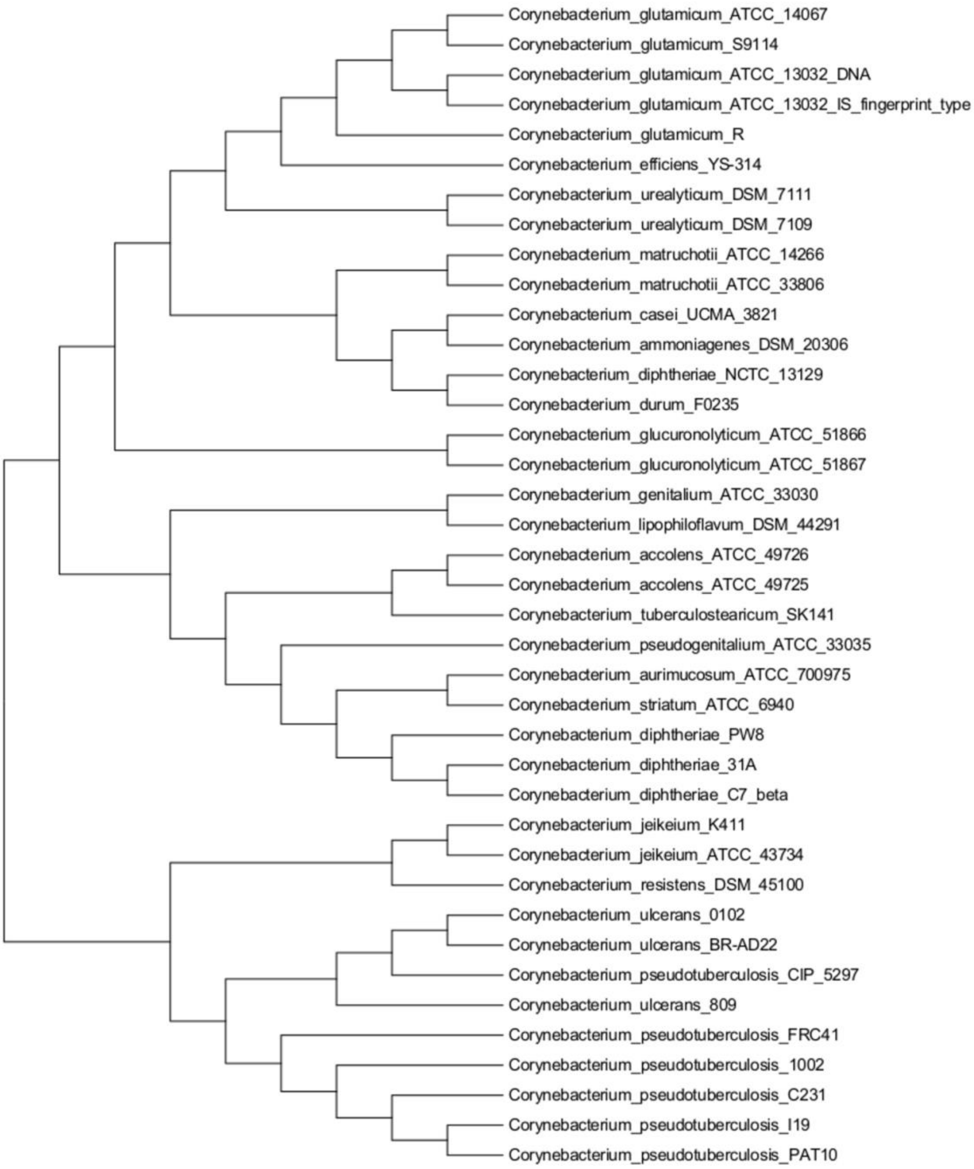
Corynebacterium_glutamicum_ATCC_14067
Corynebacterium_glutamicum_S9114
Corynebacterium_glutamicum_ATCC_13032_DNA
Corynebacterium_glutamicum_ATCC_13032_IS_fingerprint_type
Corynebacterium_glutamicum_R
Corynebacterium_efficiens_YS-314
Corynebacterium_urealyticum_DSM_7111
Corynebacterium_urealyticum_DSM_7109
Corynebacterium_matruchotii_ATCC_14266
Corynebacterium_matruchotii_ATCC_33806
Corynebacterium_casei_UCMA_3821
Corynebacterium_ammoniagenes_DSM_20306
Corynebacterium_diphtheriae_NCTC_13129
Corynebacterium_durum_F0235
Corynebacterium_glucuronolyticum_ATCC_51866
Corynebacterium_glucuronolyticum_ATCC_51867
Corynebacterium_genitalium_ATCC_33030
Corynebacterium_lipophiloflavum_DSM_44291
Corynebacterium_accolens_ATCC_49726
Corynebacterium_accolens_ATCC_49725
Corynebacterium_tuberculostearicum_SK141
Corynebacterium_pseudogenitalium_ATCC_33035
Corynebacterium_aurimucosum_ATCC_700975
Corynebacterium_striatum_ATCC_6940
Corynebacterium_diphtheriae_PW8
Corynebacterium_diphtheriae_31A
Corynebacterium_diphtheriae_C7_beta
Corynebacterium_jeikeium_K411
Corynebacterium_jeikeium_ATCC_43734
Corynebacterium_resistens_DSM_45100
Corynebacterium_ulcerans_0102
Corynebacterium_ulcerans_BR-AD22
Corynebacterium_pseudotuberculosis_CIP_5297
Corynebacterium_ulcerans_809
Corynebacterium_pseudotuberculosis_FRC41
Corynebacterium_pseudotuberculosis_1002
Corynebacterium_pseudotuberculosis_C231
Corynebacterium_pseudotuberculosis_I19
Corynebacterium_pseudotuberculosis_PAT10

Fig. 10: Phylogenetic tree based on porin genes. The construction was obtained by the Maximum Likelihood statistical method. The tree was derived from the alignments of *porA* and *porH* gene sequences. The phylogenetic distances were calculated by the software MEGA 5.

Discussion

Pan-genome, core genome, and singleton analyses

The total pan-genome of the 44 species of *Corynebacterium* genus showed an α value of 0.59. As explained before, α<1 is representative of an open pan-genome. Therefore, we may state that the pan-genome of the 44 species of *Corynebacterium* genus is still open, meaning that new genes will still be added to the pan-genome for each newly sequenced genome. Moreover, the α value is also inversely proportional to the degree of variability of the dataset, i.e., the lower the α value, the higher the variability inside the dataset. The *α* value of the 44 species of *Corynebacterium* genus is far lower than the one obtained for the pan-genome of the clonal-like species *C. pseudotuberculosis* (α=0.89) [10] and also for the highly variable species *C. diphtheriae* (α=0.69) [21]. This low *α* value is, then, indicative of the high variability of the 44 species. This prospect of the pan-genome being highly variable may also be suggested from the high number of non-redundant genes present in this dataset (22,177), which is 7.97-fold and 4.63-fold the size of the pan-genome subsets of *C. pseudotuberculosis* and *C. diphtheriae*, respectively [10, 21]. Although previous reports have indicated high conservation and synteny between the *Corynebacterium* species [2], the variability found with the pan-genomic analyses performed herein considering the whole genus was expectedly high. The explanation underlying this apparent inconsistency is that the genus comprises a large number of species with different lifestyles. Therefore, they are expected to have adapted their gene content to their particular host and environment, each of them contributing a large number of genes (singletons) to the pan-genome subset. Indeed, this scenario is also indicated in the core genome subset and in the singleton extrapolation of the *Corynebacterium* genus.

The core genome of the 44 species is composed of 562 genes, whereas in pan-genomic analyses of *C. diphtheriae* and *C. pseudotuberculosis,* the core genomes contained 1632 and 1504 genes, respectively [10, 21]. The singleton extrapolation showed that ~236 genes are expected to be added to the pan-genome of the 44 species for each newly sequenced genome, whereas in pan-genomic analyses of *C. pseudotuberculosis* and *C. diphtheriae*, the total number of singletons expected to be added to their specific pan-genomes is ~19 and ~65 genes, respectively [10, 21]. Because our dataset is composed of bacteria with different lifestyles, the low number of genes of the core genome subset and the high number of singletons predicted to be added to the *Corynebacterium* pan-genome suggests the conservation of genes with basic cellular functions and the acquisition of genes that may confer adaptative advantage to new environments.

Differential core genome analyses: pathogenic, non-pathogenic, and opportunistic

The core genome for all species showed a large amount of genes in the “Metabolism” category followed by “Information Storage and Processing” category. The first category contains genes involved in energy production and conversion, as well as transport and metabolism of carbohydrates, amino acids, nucleotides, and others. The second category contains genes involved in translation, ribosomal structure and biogenesis, RNA processing and modification, transcription, replication, recombination and repair, and other important functions. This result reflects the high importance of these two categories, containing genes involved in essential biochemical pathways [22]. This high proportion of genes in the categories “Metabolism” and “Information Storage and Processing” is also in agreement with previous pan-genomic studies using *C. pseudotuberculosis* and *Aggregatibacter actinomycetemcomitans* [23]. On the other hand, the low proportion of genes in the “Poorly Characterized” category is in disagreement with the abovementioned studies. However, genus-level analyses assume that only highly essential genes are present in the core genome, which possibly explains the small number of “Poorly Characterized” genes.

The same assumption underlies comparisons between differential core genomes with the total core genome. In the differential core genomes, a higher number of genes are related to “Metabolism” and “Poorly Characterized” genes than the total core genome; conversely, there is a lower number in “Information Storage and Processing.” Thus, it is likely that the differential core genome contains genes that may be associated to the specific environments and hosts of the species in the given subsets.

Phylogenetic analyses based in different datasets

Phylogenetic trees to correlate *Corynebacterium* species studied with their lifestyles were performed using five different strategies: *16S* rRNA, *rpoB*, whole-genome, two porin genes, and five conserved proteins.

The pathogenic species *C. pseudotuberculosis*, *C. ulcerans*, and *C. diphtheriae* formed a closely related cluster. Likewise, the non-pathogenic species *C. efficiens* and *C. glutamicum* also formed a closely related cluster. However, the phylogenomic tree based on *16S* rRNA positioned *C. diphtheriae* NCTC 13129 and *C. glutamicum* S9114 in different clusters. Although previous studies have shown that both *rpoB* and *16S* rRNA are the genes of choice for phylogenetic analyses of *Corynebacterium* genus [9, 24], our results highlight that *rpoB* gene analyses present better accuracy than *16S* rRNA and, therefore, should be adopted as the method of choice for single-gene phylogenetic analyses of this genus.

The pathogenic species *C. kroppenstedtii* and the opportunistic pathogenic species *C. amycolatum* were closely related in all phylogenomic strategies. Interestingly, only three species of *Corynebacterium* are characterized by the lack of mycolic acids in the cell wall: *C. atypicum*, *C. amycolatum*, and *C. kroppenstedtii*, which is in agreement with their close relationship in all phylogenomic trees [25]. The pathogenic species *C. bovis* and the non-pathogenic species *C. variabile* and *C. nuruki* were closely related in all strategies. Interestingly, these three species share in common the ability to use glucose, fructose, mannose, and ribose as carbon and energy sources [5, 7, 26]. The non-pathogenic species *C. casei* and the opportunistic species *C. ammoniagenes* cluster together in all phylogenetic trees, which is in agreement with the previously described high similarity (98%) between these two species [27].

Finally, the opportunistic species *C. genitalium*, *C. lipophiloflavum*, *C. aurimucosum*, *C. striatum*, *C. accolens*, *C. pseudogenitalium*, and *C. tuberculostearicum* are closely related in all phylogenetic trees. Besides, the opportunistic species *C. jeikeium*, *C. urealyticum*, and *C. resistens* tend to always cluster together, except for the concatenated two porins tree, where *C. urealyticum* presents a close relationship with the non-pathogenic species *C. glutamicum* and *C. efficiens*.

Altogether, we suggest that the *Corynebacterium* genus is composed of four major clusters based on lifestyle: one pathogenic cluster, with *C. pseudotuberculosis*, *C. diphtheriae*, and *C. ulcerans*; one non-pathogenic cluster, with *C. glutamicum* and *C. efficiens*; and two opportunistic clusters, one represented by *C. jeikeium*, *C. urealyticum*, and *C. resistens*, and the other by *C. genitalium*, *C. lipophiloflavum*, *C. aurimucosum*, *C. striatum*, *C. accolens*, *C. pseudogenitalium*, and *C. tuberculostearicum*.

Conclusions

At the *Corynebacterium* genus level, this is the first research that works with 25 different species for better understanding and evaluating the evolutionary relationship between pathogenic, non-pathogenic, and opportunistic pathogenic species.

The pan-genome extrapolation shows that *Corynebacterium* genus has an open pan-genome and new genes will be added for each newly sequenced genome. Comparing the core genome and singleton results with other studies, we found a low number for core genes and a high number for singletons, which mainly comes from the fact that our study considered bacteria with different lifestyles. The core genome for all species showed a large number of genes in “Metabolism” followed by “Information Storage and Processing”; this reflects the high importance of these two categories, containing genes involved in essential biochemical pathways. On the other hand, the differential core genome analyses for pathogenic, non-pathogenic, and opportunistic species showed a higher number of genes related to “Metabolism” and “Poorly Characterized” according to COG classification; here, we can infer that these genes are involved in the adaptation to the specific environment and hosts. The phylogenomic analysis for the corynebacteria revealed that the species of this genus are divided among four clusters based on their lifestyle: one pathogenic cluster, one non-pathogenic cluster, and two opportunistic pathogenic clusters.

This way, the comparative genomic study of the *Corynebacterium* genus shows a foundation for identification and analysis of new genes that may be involved in more diverse adaptation to different hosts and environments.

Methods

Corynebacterial genome sequences

All 44 genome sequences of the genus *Corynebacterium* used in this work are publicly available from the GenBank database. Accession numbers, the status of the genome sequence (complete or draft), the origin of the sequenced strain, and additional information are summarized in Table 1. The species were classified as pathogenic (*C. diphtheriae*, *C. pseudotuberculosis*, and *C. ulcerans*) [10, 21, 28]; opportunistic pathogenic (*C. accolens*, *C. ammoniagenes*, *C. amycolatum*, *C. aurimucosum*, *C. bovis*, *C. durum*, *C. genitalium*, *C. glucuronolyticum*, *C. kroppenstedtii*, *C. jeikeium*, *C. lipophiloflavum*, *C. matruchotii*, *C. pseudogenitalium*, *C. resistens*, *C. striatum*, *C. tuberculostearicum*, and *C. urealyticum*) [29–38]; and non-pathogenic (*C. casei*, *C. efficiens*, *C. glutamicum*, *C. nuruki*, and *C. variabile*) [29, 38–41]. There is not a consensus if the species *C. bovis* and *C. kroppenstedtii* are pathogenic or opportunistic pathogens species [32, 37]. In this work, we considered both as opportunistic pathogens.

Pan-genome, core genome and singleton analysis

Genomes were compared using the EDGAR software (version 1.2) which performs homology analyses based on specific cutoffs automatically adjusted to the query data [16]. Additionally, for the core genome analysis the genomes were divided into: (i) pathogens where *C. pseudotuberculosis* strain FRC41 was used as reference; (ii) opportunistic pathogens where *C. urealyticum* strain DSM 7109 was used as reference; and (iii) non-pathogens where *C. glutamicum* strain R was used as reference.

Orthology analysis to calculate the sizes of the pan-genome and the core genome and the number of singletons was performed using BLAST Score Ratio Values (SRV). This method divides the BLAST bit score by the maximum possible bit score, thereby generating the SRV. The cutoff was calculated by the EDGAR software using a sliding window, instead of a fixed SRV as proposed previously [42].

The core genome was predicted through iterative pairwise comparison using all 44 genomes. One genome was selected as a reference and its gene set (A) was compared with another gene set (B). Genes with a reciprocal best BLAST hit (A and B gene set) were filtered according to the orthology criterion based on the SRV, and this new gene subset formed the core AB. Subsequently, this subset was compared with another gene set (C), and these comparisons were continued with all 44 genomes. The corynebacterial pan-genome was predicted in the same way, however, adding also non-orthologous genes. Singletons were predicted as genes present in only one corynebacterial genome. Extrapolations of the pan-genome, the core genome and the detected singletons were calculated based on 10,000 randomly selected permutations of the set of 44 genomes. The pan-genome extrapolation was performed using Heaps' Law ($n = \kappa \times N^{\gamma}$) estimating the empiric parameters κ and γ. The core genome and singleton extrapolations were calculated using the least-squares fit of the exponential regression decay to the mean values [12, 13].

COG (Cluster of Orthologous Genes) was used to classify the proteins inside the core genome as: (i) Information storage and processing; (ii) Cellular processes and signaling; (iii) Metabolism; and (iv) Poorly characterized. Additionally, the differential core genomes of pathogen, non-pathogen, and opportunistic pathogen species were also submitted to the COG database and a BLAST protein was performed to look for similarity among proteins. All proteins that showed E-values higher than 10^{-6} were excluded, and the best BLAST results for each protein were considered for the COG classification.

Phylogenomic analysis

The phylogenomic analyses were done to compare different strategies and resolutions of trees combining whole genome-based, single-gene, and gene concatenation.

For the whole genome-based analysis, the generation of a phylogenomic tree was based on the core genome predicted by the EDGAR software (version 1.2) [16]. All orthologous genes predicted for the 44 genomes were used to build a multiple protein alignment using MUSCLE [43]. The matching parts of the alignments were concatenated, whereas the non-matching ones were excluded using GBLOCKS [44]. The data were exported from EDGAR as a distance matrix file in nexus format and were imported into the MEGA software [19] to obtain a phylogenomic tree. In addition, Gegenees (version 2.0) was used to generate a heatmap chart [45]. For this analysis, the software divided the genome sequences into small pieces and performed a BLAST search for similarity to define the common sequence content shared by all genomes. The common sequence content was compared against all genomes to generate the similarity percentages which were used to plot the heatmap chart.

Two phylogenetic trees were generated for the single genes analyses using *16S* rRNA and *rpoB* genes (β-subunit of RNA polymerase). *16S* rRNA was chosen because this gene is extremely informative for phylogenomic analysis [46]. *rpoB* gene was chosen because this gene showed a good accuracy for the identification of *Corynebacterium* species [9, 24]. The sequences were separately aligned using MUSCLE [43]. After that, we used the Maximum Likelihood statistical method to generate the phylogenomic trees in MEGA.

Analysis using gene concatenation was performed for two subsets of conserved genes: 5 proteins (*isoleucyl tRNA synthetase, ribosomal protein S1*, *DNA topoisomerase, SecY*, and *GTPase*) involved in conserved cellular functions [20]; and *porA* and *porH*, which have been shown to be associated with host-cell recognition and invasiveness during the initial stage of infection, thereby contributing to pathogenicity and virulence, as well as, allow permeation of small hydrophilic molecules across the outer membrane permeability barrier [47–49]. The software MUSCLE was used [43] in both gene sets to create a multiple alignment and, after that, we used the Maximum Likelihood statistical method to generate the phylogenomic trees in MEGA.

Competing interests

The authors declare that there are not competing interests.

Authors' contributions

Read and gave insights about the manuscript: LCG, CB, SCS, MPS, RTJR, MPCS, AM, AS, VA, AT. Conceived and designed the experiments: VA, AT. Executed the experiments: LCG, JB, SCS, RTJR, PHCGS, UPP. Analyzed the data: LCG, CB, SCS, MPS, TRJR, PHCGS, UPP. Subsidized reagents/ materials/ analyses tools: JB, SCS, MPCS, AM, AS, VA, AT. Wrote the paper: LCG, CB, SCS, MPS, MPCS, AS, AM, VA, AT.

ACKNOWLEDGMENTS

The authors thank the Coordination of Improvement of Higher Education Personnel/ Doctoral Sandwich Program Exchange (CAPES/PDSE) international cooperation for financing an exchange scholarship (grant number 9151125) and the CLIB Graduate Cluster Industrial Biotechnology at Bielefeld University (Bielefeld, Germany), where LCG participated as a guest student.

REFERENCES

1. Pagani I, Liolios K, Jansson J, Chen I-M a, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC: **The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata.** *Nucleic Acids Res* 2012, **40**(Database issue):D571–9.

2. Ventura M, Canchaya C, Tauch A, Chandra G, Fitzgerald GF, Chater KF, van Sinderen D: **Genomics of Actinobacteria: tracing the evolutionary history of an ancient phylum.** *Microbiol Mol Biol Rev* 2007, **71**:495–548.

3. Dorella FA, Pacheco LGC, Oliveira SC, Miyoshi A, Azevedo V: **Corynebacterium pseudotuberculosis : microbiology , biochemical properties , pathogenesis and molecular studies of virulence**. 2006, **37**:201–218.

4. Cerdeno-Tarraga a. M: **The complete genome sequence and analysis of Corynebacterium diphtheriae NCTC13129**. *Nucleic Acids Res* 2003, **31**:6516–6523.

5. Schröder J, Maus I, Meyer K, Wördemann S, Blom J, Jaenicke S, Schneider J, Trost E, Tauch A: **Complete genome sequence, lifestyle, and multi-drug resistance of the human pathogen Corynebacterium resistens DSM 45100 isolated from blood samples of a leukemia patient.** *BMC Genomics* 2012, **13**:141.

6. Tauch A, Trost E, Tilker A, Ludewig U, Schneiker S, Goesmann A, Arnold W, Bekel T, Brinkrolf K, Brune I, Götker S, Kalinowski J, Kamp P-B, Lobo FP, Viehoever P, Weisshaar B, Soriano F, Dröge M, Pühler A: **The lifestyle of Corynebacterium urealyticum derived from its complete genome sequence established by pyrosequencing.** *J Biotechnol* 2008, **136**:11–21.

7. Schröder J, Maus I, Trost E, Tauch A: **Complete genome sequence of Corynebacterium variabile DSM 44702 isolated from the surface of smear-ripened cheeses and insights into cheese ripening and flavor generation.** *BMC Genomics* 2011, **12**:545.

8. Tauch A, Kaiser O, Hain T, Goesmann A, Weisshaar B, Albersmeier A, Bekel T, Bischoff N, Brune I, Chakraborty T, Meyer F, Rupp O, Schneiker S, Viehoever P, Pu A: **Complete Genome Sequence and Analysis of the Multiresistant Nosocomial Pathogen Corynebacterium jeikeium K411 , a Lipid-Requiring Bacterium of the Human Skin Flora**. 2005, **187**:4671–4682.

9. Khamis A, Raoult D, Scola B La: **rpoB Gene Sequencing for Identification of Corynebacterium Species rpoB Gene Sequencing for Identification of Corynebacterium Species**. 2004, **42**.

10. Soares SC, Silva A, Trost E, Blom J, Ramos R, Carneiro A, Ali A, Santos AR, Pinto AC, Diniz C, Barbosa EG V, Dorella F a, Aburjaile F, Rocha FS, Nascimento KKF, Guimarães LC, Almeida S, Hassan SS, Bakhtiar SM, Pereira UP, Abreu V a C, Schneider MPC, Miyoshi A, Tauch A, Azevedo V: **The Pan-Genome of the Animal Pathogen Corynebacterium pseudotuberculosis Reveals Differences in Genome Plasticity between the Biovar ovis and equi Strains.** *PLoS One* 2013, **8**:e53818.

11. Pallen MJ, Wren BW: **Bacterial pathogenomics.** *Nature* 2007, **449**:835–42.

12. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli S V, Crabtree J, Jones AL, Durkin a S, Deboy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan S a, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, et al.: **Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome".** *Proc Natl Acad Sci U S A* 2005, **102**:13950–5.

13. Tettelin H, Riley D, Cattuto C, Medini D: **Comparative genomics: the bacterial pan-genome**. *Curr Opin Microbiol* 2008, **11**:472–477.

14. Mira A, Martín-Cuadrado AB, D'Auria G, Rodríguez-Valera F: **The bacterial pan-genome:a new paradigm in microbiology**. *Int Microbiol* 2010, **13**:45—57.

15. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R: **The microbial pan-genome.** *Curr Opin Genet Dev* 2005, **15**:589–94.

16. Blom J, Albaum SP, Doppmeier D, Pühler A, Vorhölter F-J, Zakrzewski M, Goesmann A: **EDGAR: a software framework for the comparative analysis of prokaryotic genomes.** *BMC Bioinformatics* 2009, **10**:154.

17. Suzuki N, Okai N, Nonaka H, Tsuge Y, Inui M, Yukawa H: **High-Throughput Transposon Mutagenesis of Corynebacterium glutamicum and Construction of a Single-Gene Disruptant Mutant Library High-Throughput Transposon Mutagenesis of Corynebacterium glutamicum and Construction of a Single-Gene Disruptant Mutant Libra**. 2006.

18. Ventura M, Canchaya C, Tauch A, Chandra G, Fitzgerald GF, Chater KF, van Sinderen D: **Genomics of Actinobacteria: tracing the evolutionary history of an ancient phylum.** *Microbiol Mol Biol Rev* 2007, **71**:495–548.

19. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.** *Mol Biol Evol* 2011, **28**:2731–9.

20. Alam MT, Merlo ME, Takano E, Breitling R: **Genome-based phylogenetic analysis of Streptomyces and its relatives.** *Mol Phylogenet Evol* 2010, **54**:763–72.

21. Trost E, Blom J, Soares SDC, Huang I-H, Al-Dilaimi A, Schröder J, Jaenicke S, Dorella F a, Rocha FS, Miyoshi A, Azevedo V, Schneider MP, Silva A, Camello TC, Sabbadini PS, Santos CS, Santos LS, Hirata R, Mattos-Guaraldi AL, Efstratiou A, Schmitt MP, Ton-That H, Tauch A: **Pangenomic study of Corynebacterium diphtheriae that provides insights into the genomic diversity of pathogenic isolates from cases of classical diphtheria, endocarditis, and pneumonia.** *J Bacteriol* 2012, **194**:3199–215.

22. Tatusov RL, Galperin MY, Natale DA, Koonin E V: **The COG database : a tool for genome-scale analysis of protein functions and evolution**. 2000, **28**:33–36.

23. Kittichotirat W, Bumgarner RE, Asikainen S, Chen C: **Identification of the pangenome and its components in 14 distinct Aggregatibacter actinomycetemcomitans strains by comparative genomic analysis.** *PLoS One* 2011, **6**:e22420.

24. Khamis A, Raoult D, Scola B La: **Comparison between rpoB and 16S rRNA Gene Sequencing for Molecular Identification of 168 Clinical Isolates of Corynebacterium Comparison between rpoB and 16S rRNA Gene Sequencing for Molecular Identification of 168 Clinical Isolates of Corynebacterium**. 2005.

25. Collins MD, Burton RA, Jones D: **Corynebacterium amycolatum sp. nov. a new mycolic acid-less Corynebacterium species from human skin**. *FEMS Microbiol Lett* 1988, **49**:349–352.

26. Shin N-R, Whon TW, Roh SW, Kim M-S, Jung M-J, Lee J, Bae J-W: **Genome sequence of Corynebacterium nuruki S6-4 T, isolated from alcohol fermentation starter.** *J Bacteriol* 2011, **193**:4257.

27. George K, Building VI, Brennan NM, Brown R, Goodfellow M, Ward AC, Beresford TP, Simpson PJ, Fox PF, Cogan TM: **Corynebacterium casei sp . nov ., isolated from the surface of a smear-ripened cheese**. 2001:843–852.

28. Trost E, Al-Dilaimi A, Papavasiliou P, Schneider J, Viehoever P, Burkovski A, Soares SC, Almeida SS, Dorella F a, Miyoshi A, Azevedo V, Schneider MP, Silva A, Santos CS, Santos LS, Sabbadini P, Dias A a, Hirata R, Mattos-Guaraldi AL, Tauch A: **Comparative analysis of two complete Corynebacterium ulcerans genomes and detection of candidate virulence factors.** *BMC Genomics* 2011, **12**:383.

29. Pindi PK, Yadav PR, Shanker AS: **Identification of Opportunistic Pathogenic Bacteria in Drinking Water Samples of Different Rural Health Centers and Their Clinical Impacts on Humans**. 2013, **2013**.

30. Cooke J V, Keith HR: **A Type Of Urea-Splitting Bacterium Found In The Human Intestinal Tract**. 1926, **XIII**:315–319.

31. Trost E, Götker S, Schneider J, Schneiker-Bekel S, Szczepanowski R, Tilker A, Viehoever P, Arnold W, Bekel T, Blom J, Gartemann K-H, Linke B, Goesmann A, Pühler A, Shukla SK, Tauch A: **Complete genome sequence and lifestyle of black-pigmented Corynebacterium aurimucosum ATCC 700975 (formerly C. nigricans CN-1) isolated from a vaginal swab of a woman with spontaneous abortion.** *BMC Genomics* 2010, **11**:91.

32. Honkanen-Buzalski T, Griffin TK, Dodd FH: **Observations on Corynebacterium bovis infection of the bovine mammary gland: I. Natural infection**. *J Dairy Res* 1984, **51**:371–378.

33. Rizvi M, Khan F, Raza A, Shukla I, Malik A, Ali S, Rizvi R, Sherwani MK, Afzal K, Hasan SA: **Coryneforms the Opportunistic Pathogens - An Emerging Challenge for Immunocompetent Individuals**. 2011, **6**:165–171.

34. Barrett SLR, Cookson BT, Ladonna C, Bernard KA, Coyle MB, Carlson LC: **Diversity within Reference Strains of Corynebacterium matruchotii Includes Corynebacterium durum and a Novel Organism Diversity within Reference Strains of Corynebacterium matruchotii Includes Corynebacterium durum and a Novel Organism**. 2001.

35. Riegel P, Heller R, Prevost G, Jehl F, Monteil H: **Corynebacterium durum sp . nov ., from Human Clinical SDecimens**. 1997:1107–1111.

36. Devriese LUCA, Riegel P, Hommez J, Vaneechoutte M, Baere TDE, Haesebrouck F: **Identification of Corynebacterium glucuronolyticum Strains from the Urogenital Tract of Humans and Pigs**. 2000, **38**:4657–4659.

37. Kazmierczak AK, Szarapinska-Kwaszewska JK, Szewczyk EM: **Opportunistic Coryneform Organisms – Residents of Human Skin**. 2005, **54**.

38. Tauch A, Bischoff N, Pühler A, Kalinowski J: **Comparative genomics identified two conserved DNA modules in a corynebacterial plasmid family present in clinical isolates of the opportunistic human pathogen Corynebacterium jeikeium.** *Plasmid* 2004, **52**:102–18.

39. Shin N-R, Whon TW, Roh SW, Kim M-S, Jung M-J, Lee J, Bae J-W: **Genome sequence of Corynebacterium nuruki S6-4 T, isolated from alcohol fermentation starter.** *J Bacteriol* 2011, **193**:4257.

40. Nishio Y, Nakamura Y, Kawarabayasi Y, Usuda Y, Kimura E, Sugimoto S, Matsui K, Yamagishi A, Kikuchi H, Ikeo K, Gojobori T: **Comparative Complete Genome Sequence Analysis of the Amino Acid Replacements Responsible for the Thermostability of Corynebacterium efficiens**. 2003:1572–1579.

41. Monnet C, Loux V, Bento P, Gibrat J-F, Straub C, Bonnarme P, Landaud S, Irlinger F: **Genome sequence of Corynebacterium casei UCMA 3821, isolated from a smear-ripened cheese.** *J Bacteriol* 2012, **194**:738–9.

42. Lerat E, Daubin V, Moran N a: **From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria.** *PLoS Biol* 2003, **1**:E19.

43. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32**:1792–7.

44. Talavera G, Castresana J: **Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments.** *Syst Biol* 2007, **56**:564–77.

45. Agren J, Sundström A, Håfström T, Segerman B: **Gegenees: fragmented alignment of multiple genomes for determining phylogenomic distances and genetic signatures unique for specified target groups.** *PLoS One* 2012, **7**:e39107.

46. Woese CR: **Bacterial Evolution Background**. 1987, **51**:221–271.

47. Hünten P, Costa-Riu N, Palm D, Lottspeich F, Benz R: **Identification and characterization of PorH, a new cell wall channel of Corynebacterium glutamicum.** *Biochim Biophys Acta* 2005, **1715**:25–36.

48. Costa-riu N, Burkovski A, Krämer R, Benz R, Kra R: **PorA Represents the Major Cell Wall Channel of the Gram-Positive Bacterium Corynebacterium glutamicum PorA Represents the Major Cell Wall Channel of the Gram-Positive Bacterium Corynebacterium glutamicum**. 2003.

49. Rath P, Demange P, Saurel O, Tropis M, Daffé M, Dötsch V, Ghazi A, Bernhard F, Milon A: **Functional expression of the PorAH channel from Corynebacterium glutamicum in cell-free expression systems: implications for the role of the naturally occurring mycolic acid modification.** *J Biol Chem* 2011, **286**:32525–32.

IV.4. Chapter V *Corynebacterium ulcerans*

This chapter is divided in six sections. The first section is a brief introduction on the organism a and disease caused by *Corynebacterium ulcerans*, the second section refer to genomic data available and, third to sixth sections treat about the genome announcements of the *Corynebacterium ulcerans* strains 210932, 210931, FRC11 and, 05146 respectively.

IV.4.1. *Corynebacterium ulcerans* – State of the art

The first *Corynebacterium ulcerans* isolation occured in 1926 in human throat lesions, but only in 1995 it was recognised as a distinct species (GILBERT; STEWART, 1926; RIEGEL *et al.*, 1995). Isolation of *C. ulcerans* strains from many wild and domestic animals have been reported (HOGG *et al.*, 2008; LARTIGUE *et al.*, 2005). In some cases, it has been hypothesized that infections in humans are derived from pets that are colonized by this bacterial species (ZOYSA, DE *et al.*, 2005) highlighting the zoonotic potential of this organism. In humans, *C. ulcerans* causes diphtheria-like disease associated with pharyngitis, sinusitis, tonsillitis, pulmonary nodules, and skin ulcers (BERNARD, 2012; KAUFMANN *et al.*, 2002; WAGNER *et al.*, 2012; WELLINGHAUSEN *et al.*, 2002). It is considered an emergent pathogen for two reasons: (i) the number of cases of infection in humans has been constantly increasing around the world and (ii) recently toxigenic strains have been described encoding the tox gene causing diphtheria-like disease (DIAS *et al.*, 2011).

Analyses using the *rpoB* gene (β subunit of RNA polymerase) revealed that *C. ulcerans* is closely related with other important pathogens from the *Corynebacterium* genus such as *C. diphteriae* and *C. pseudotuberculosis* (Figure 1) (KHAMIS *et al.*, 2004, 2005).

Some *C. ulcerans* strains encode the *tox* gene similar to that described in the pathogenic species *C. diphteriae* (SEKIZUKA *et al.*, 2012; SING *et al.*, 2004). This *tox* gene, which encodes the diphtheria toxin (DT), is present in a lysogenic β-corynephage that was described in *C. ulcerans* 0102 and other strains but not in 809 strain, which, instead, contains a remnant phage-related integrase (SEKIZUKA *et al.*, 2012). Phopholipase D (PLD) normally produced by *C. pseudotuberculosis* was also described in *C. ulcerans* (TROST *et al.*, 2011). This exotoxin facilitates the permeability promoting the hydrolysis of ester bonds in sphingomyelin in mammalian cell membranes, probably contributing to the spread of the *C. pseudotuberculosis* (DORELLA *et al.*, 2006). A research using 10 *C. ulcerans* strains showed that the majority of the isolates encoded both exotoxins (CARNE; ONON, 1982). However, there are more virulence factors encoded by *C. ulcerans* as, neuraminidase H, endoglycosidase EndoE, and a ribosome-binding protein with structural similarity to Shiga-like toxins (TROST *et al.*, 2011).

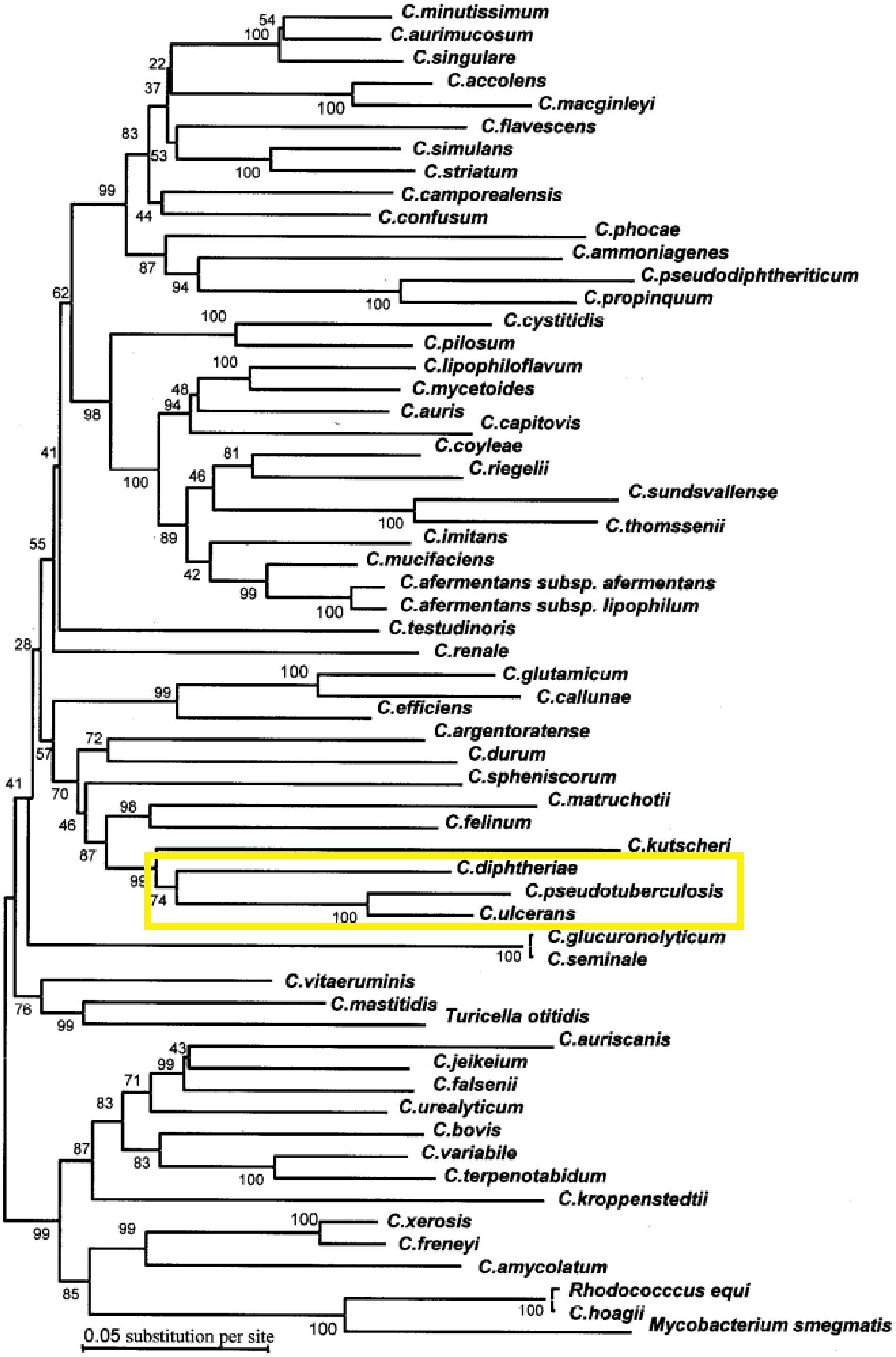

Figure 1. Phylogenetic tree based on *rpoB* gene representing the genetic relationships of *Corynebacterium* genus. Adapted from Khamis et al (2004).

Even *C. ulcerans* being an important pathogen, its pathogenic mechanism is not yet fully understood. Therefore, it is essential to perform a more detailed analysis of genomic features of this bacterial species as done for *C. pseudotuberculosis* (SOARES *et al.*, 2013) and, *C. diphteriae* (TROST *et al.*, 2012) closely related pathogenic organisms, as showed in figure 1. These genomic studies were developed in a cooperation network between Universidade Federal de Minas Gerais (Belo Horizonte, Brazil) and Universidade Federal do Pará (Belém, Brazil). From this cooperation network, 15 strains of the animal pathogen *C. pseudotuberculosis* were sequenced, assembled, annotated and deposited at GenBank. The results of this study provided new insights about the lifestyle and molecular biology of the *Corynebacterium* genus. Additionally, a network of collaboration between the University of Bielefeld (Bielefeld, Germany), Universidade Federal de Minas Gerais (Belo Horizonte, Brazil) and Universidade Federal do Pará (Belém, Brazil) resulted in the realization of pan-genomic study of *C. diphtheriae*.

In this context, the network between Universidade Federal de Minas Gerais (Belo Horizonte, Brazil) and Universidade Federal do Pará (Belém, Brazil) have the goal to analyze the pan-genome of *C. ulcerans* species, an important human and animal pathogen with great medical interest, along with its underlying features.

IV.4.2. *Corynebacterium ulcerans data*

Currently, nine *C. ulcerans* genomes are available at GenBank, of which seven are completely closed and two are drafts (Table 1). The strains 210932, 210931, FCR11, 05146 and, FRC58 were sequenced, assembled and, deposited through cooperation network between Universidade Federal de Minas Gerais (Belo Horizonte, Brazil) and Universidade Federal do Pará (Belém, Brazil). Additionally, seven more strains were sequenced and their status may be checked at Table 2. These genomic data were obtained from different platforms of next generation sequencing: SOLiD 5500 System (Apllied Biosystems) and Ion Torrent Personal Genome Machine System (Life Sciences).

Table 1. General information about the 9 *C. ulcerans* strains available at GenBank

Strain	**Genome Size**	**CG %**	**Number of genes**	**GenBank accession number**	**Status**	**Reference**
BR-AD22	2.60637	53.40	2402	CP002791.1	complete	(TROST *et al.*, 2011)
809	2.5021	53.30	2246	CP002790.1	complete	(TROST *et al.*, 2011)
0102	2.57919	53.40	2417	AP012284.1	complete	(SEKIZUKA *et al.*, 2012)
210932	2.48434	53.30	2344	CP009500.1	complete	(VIANA *et al.*, 2014)
210931	2.50943	53.25	2303	CP009583.1*	complete	-
FRC11	2.44283	53.30	2210	CP009622.1	complete	-
05146	2.46644	53.30	2336	CP009716.1	complete	-
NCTC12077	2.61629	53.40	2457	AYUJ00000000.1	draft	-
FRC58	2.60941	53.20	2430	AYTI00000000.1	draft	(SILVA *et al.*, 2014)

* Plasmid GenBank accession number: CP009584.1.

Table 2. Status information about the 7 *C.ulcerans* strains sequenced by our network collaboration

Strain	Assembly	Annotation
131001	Finished	Mannual curation
131002	Finished	Mannual curation
03_8664	GAPs closure - 119 contigs	Waiting finish assembly
03_9258	GAPs closure - 375 contigs	Waiting finish assembly
04_3911	GAPs closure - 44 contigs	Waiting finish assembly
04_7514	GAPs closure - 7 contigs	Waiting finish assembly
CIP54.53	GAPs closure - 5 contigs	Waiting finish assembly

All strains showed in table 2 will be deposited at GenBank and their respective genome announcements will be available in the literature.

Subsequently, a genome comparison will be done using all sixteen genomes, once it is essential to achieve a better detailing about this bacterial species. With the large amount of *C. ulcerans* sequenced data, it becomes possible to analyze its core genome, pathogenicity islands, phylogenomic relationships and other important characteristics.

Next sections are dedicated to genome announcements of the strains 210932, 210931, FRC11 and 05146.

IV.4.3. Genome Sequence of *Corynebacterium ulcerans* strain 210932[5]

Marcus Vinicius Canário Viana, Leandro de Jesus Benevides, Diego Cesar Batista Mariano, Flávia de Souza Rocha, Priscilla Carolinne Bagano Vilas Boas, Edson Luiz Folador, Felipe Luiz Pereira, Fernanda Alves Dorella, Carlos Augusto Gomes Leal, Alex Fiorini de Carvalho, Artur Silva, Siomar de Castro Soares, Henrique Cesar Pereira Figueiredo, Vasco Azevedo, **Luis Carlos Guimarães**

In this section, we present an article announcing the sequencing, assembly, annotation and deposit at GenBank of the *Corynebacterium ulcerans 210932*, isolated from human. Its genome has one circular chromosome with 2,484,335 bp encoding 2,282 genes, 12 rRNAs and 51 tRNAs. The genome project is complete; under accession number CP009500.1.

[5] Published article in Genome Announcements.

IV.4.4. The complete genome sequence of *Corynebacterium ulcerans* strain 210931[6]

Leandro de Jesus Benevides, Marcus Vinicius Canário Viana, Diego César Batista Mariano, Flávia de Souza Rocha, Priscilla Carolinne Bagano, Edson Luiz Folador, Felipe Luiz Pereira, Fernanda Alves Dorella, Carlos Augusto Gomes Leal, Alex Fiorini de Carvalho, Artur Silva, Siomar de Castro Soares, Henrique César Pereira Figueiredo, Vasco Azevedo, **Luis Carlos Guimarães**

In this section, we present an article announcing the sequencing, assembly, annotation and deposit at GenBank of the *Corynebacterium ulcerans 210931*, isolated from human. The genome project is complete; under accession number CP009583.1. Additionally, for this strain, we found the plasmid PML21, 5.708pb longer, which is also present in *Enterococcus faecalis.* The plasmid was assembled, annotated and deposited at GenBank under accession number CP009584.1.

[6] Submitted article to SIGS in 15-oct-2014.

IV.4.5. The complete genome sequence of *Corynebacterium ulcerans* strain FRC11[7]

Leandro de Jesus Benevides, Marcus Vinicius Canário Viana, Diego César Batista Mariano, Flávia de Souza Rocha, Priscilla Carolinne Bagano, Edson Luiz Folador, Felipe Luiz Pereira, Fernanda Alves Dorella, Carlos Augusto Gomes Leal, Alex Fiorini de Carvalho, Siomar de Castro Soares, Adriana Carneiro, Rommel Ramos, Edgar Badell-Ocando, Nicole Guiso, Artur Silva, Henrique Figueiredo, Vasco Azevedo, **Luis Carlos Guimarães**.

In this section, we present an article announcing the sequencing, assembly, annotation and deposit at GenBank of the *Corynebacterium ulcerans* FRC11, isolated from human. The genome is 2,442,826 bp in size, encoding 2,210 genes, 12 rRNAs and 51 tRNAs The genome project is complete; under accession number CP009622.1.

[7] Published article in Genome Announcements.

IV.4.6. Genome sequence of *Corynebacterium ulcerans* strain 05146[8]

Marcus Vinicius Canário Viana, Leandro de Jesus Benevides, Diego César Batista Mariano, Flávia de Souza Rocha, Priscilla Bagano, Edson Luiz Folador, Felipe Luiz Pereira, Fernanda Alves Dorella, Carlos Augusto Gomes Leal, Alex Fiorini de Carvalho, Artur Silva, Siomar de Castro Soares, Henrique César Pereira Figueiredo, Vasco Azevedo, **Luis Carlos Guimarães**.

In this section, we present an article announcing the sequencing, assembly, annotation and deposit at GenBank of the *Corynebacterium ulcerans* 05146, isolated from human. The genome is 2,466,435 pb in size, encoding 2,322 genes, 12 rRNAs and 48 tRNAs The genome project is complete; under accession number CP009716.1.

[8] Submitted article to SIGS in 02-feb-2015

V. General Discussion

On Chapter I, we have reported the genome announcement of the *C. urealyticum* DSM 7111, isolated from the urine of a young boy with an ectopic kidney. The genome sequencing was performed using two different methodologies: (i) Pyrosequencing through 454 Genome Sequencer FLX system approach (Roche Applied Science) and (ii) semiconductor sequencing through Ion Torrent PGM Sequencer (Life Technologies). Even these two sequencing generating approximately sixty fold coverage, it was necessary to do a PCR amplification and subsequent sequencing of the DNA fragments using ABI 3730xl DNA Analyzer (Life Technologies) to close the 73 remaining GAPs. The final assembly generated one circular chromosome with 2,316,065 bp of size and an average G+C content of 64.24%. The prediction and manual annotation were performed by the GenDB platform and the REGANOR gene prediction server which revealed 1,935 protein-coding genes, 3 rRNA operons, and 54 tRNAs. A preliminary genome comparison analyses between strains DSM 7109 and DSM 7111 identified 79 exclusive genes for *C. urealyticum* DSM 7111. Some of these strain-specific genes encode a siderophore biosynthesis pathway, an iron ABC transport system, and a multicopper oxidase system, suggesting that the strains DSM 7109 and DSM 7111 have a different iron acquisition mechanism. The genome project has been deposited in GenBank under the accession number CP004085.

On Chapter II, we described a genome comparison between two *C. urealyticum* strains DSM 7109 and DSM 7111 available at GenBank. Both strains were isolated from human with alkaline-encrusted cystitis. Furthermore, these bacteria show similar genomic composition, although the strain DSM 7111 is approximately 50 kb shorter than strain DSM 7109. This genomic loss can be associated to the large number of genomic islands (26 for each genome) predicted for both genomes. This can explain the high number of multi-drug resistant genes and strain-specific genes predicted in each genome. However, 19 virulence factors were found in both strains, as *spaDEF* operon that encode an adhesive pilus responsible for facilitating the adhesion of the pathogen to host cells and with similar structure to pathogenic species like *C. diphtheriae* and *C. ulcerans*. Summarizing, this comparative genomic study of two *C. urealyticum* strains provided insights in the lifestyle of this opportunistic pathogen and highlighted some genes that can be used in a deep reverse vaccinology study to predict new antigenic targets against this bacterium.

The Chapter III discusses the pan-genome of the genus *Corynebacterium* which have pathogenic, opportunistic pathogenic and, non-pathogenic species with significant scientific interest. Moreover, this is the first research that works with 44 genomes of 25 different *Corynebacterium* species for better understanding and evaluating the evolutionary relationship between species with different lifestyles. Core genome analyses showed that these 44 genomes share 562 genes in common and core genome extrapolations showed a stability which means if more species are to be added, the core genome size will not change significantly. On the other hand, for each newly sequenced genome, ~236 genes will be added to the singletons data set. This probably happens because of the several bacteria lifestyles mentioned before. Furthermore, we can infer that the high number of singletons added by newly sequenced species is contributing to the open pan-genome predicted for this genus. Additionally, phylogenomic analysis revealed that this data set of genomes is divided in four different clusters based on their lifestyle being one pathogenic cluster, one non-pathogenic cluster, and two opportunistic pathogenic clusters.

Finally, on the chapter IV, we discussed about *C. ulcerans* and the work that is currently being performed with this species. At the end of this work we expect to obtain 16 genomes for pan-genome study. This pan-genome is important because the number of *C. ulcerans* infections has been increasing around the world and the pathogenicity mechanisms are not totally elucidated yet. Moreover, the pan-genome of *C. diphtheriae* and *C. pseudotuberculosis*, closely related organisms, were also performed by our collaboration network. Until now, we have deposited seven genomes at Genbank, from which four genomes (210932, 210931, FRC11 and, 05146) were done under my supervision. The genome announcements of these four strains are inserted at this chapter sections.

VI. Conclusions and Perspectives

Our studies in genomic era started with *C. pseudotuberculosis* and *C. diphtheriae*, two important pathogens. Now we expanded the analyses to other species of the *Corynebacterium* genus trying to understand the pathogenicity mechanisms and lifestyle diversity present in this genus. In this process, we have:

a. predicted genomic islands and new vaccine targets to *C. urealyticum*. Also, we found an important operon (*spaDEF*) present in pathogenic species which is correlated with the adhesion of the pathogen to host cells. In addition, we can infer that the number of genomic islands are correlated to multi-drug resistant status of this bacterium.
b. performed pan-genome, core genome and, singletons analyses of the *Corynebacterium* genus, along with their underlying extrapolations. Core genome extrapolations showed that if more species are to be added, the core genome size will not change significantly. This result corroborates with the results found in COG classification performed in the core genome of all species which shows a large number of genes in “Metabolism” followed by “Information Storage and Processing” categories; this represents the high importance of these genes present in the core genome once they are involved in essential biochemical pathways. Otherwise, COG classification of the differential core genome of pathogenic, non-pathogenic, and opportunistic species showed a higher number of genes related to “Metabolism” and “Poorly Characterized”. This way, we can infer that these genes are involved in the adaptation to the specific environment and hosts.
c. established efficient protocols to assemble and annotate genomes, depositing in GenBank genomes with high accuracy. An important step to pan-genomic studies.

As perpectitves to future works, we pretend to do:

a. caracterize the genes predicted in differential core-genone of the pathogenic, non-pathogenic and, opportunistic pathogenic species of the genus *Corynebacterium* with objective to find putative lifestyle related genes and to better understand the common traits of the species in the three datasets.
b. pan-genome studies of *C. ulcerans* establishing the genetic content; characterize the different features of the genomes such as replication origin and, G+C content; calculate the evolutionary history using phylogenomic techniques; identify pathogenic islands, compare the different prophages normally founded in this species and, characterize the main metabolic pathways evolved in the process of host-pathogen interaction.
c. complete study involving the three most important pathogen species of the genus *Corynebacterium* (*C. pseudotuberculosis, C. diphtheriae* and *C. ulcerans*).

VII. Bibliography

BERNARD, K. The genus corynebacterium and other medically relevant coryneform-like bacteria. **Journal of clinical microbiology**, v. 50, n. 10, p. 3152–8, out 2012.

CARNE, H. R.; ONON, E. O. The exotoxins of Corynebacterium ulcerans. **Journal of Hygiene**, v. 88, n. 02, p. 173, 25 mar 1982.

DIAS, A.; SANTOS, L.; SABBADINI, P.; *et al.* Corynebacterium ulcerans diphtheria: an emerging zoonosis in Brazil and worldwide. **Revista de Saúde …**, v. 45, n. 6, 2011.

DORELLA, F. A.; PACHECO, L. G. C.; OLIVEIRA, S. C.; MIYOSHI, A.; AZEVEDO, V. Corynebacterium pseudotuberculosis : microbiology , biochemical properties , pathogenesis and molecular studies of virulence. v. 37, p. 201–218, 2006.

GILBERT, R.; STEWART, F. C. Corynebacterium ulcerans: a pathogenic micro-organism resembling C. diphtheriae. **Journal of Laboratory and Clinical Medicine**, v. 12, p. 756–761, 1926.

HOGG, R. A.; WESSELS, J.; HART, J.; *et al.* SHORT COMMUNICATIONS Possible zoonotic transmission of toxigenic. 2008.

KAUFMANN, D.; OTT, P.; ZBINDEN, R. Laryngopharyngitis by Corynebacterium ulcerans. **Infection**, v. 30, n. 3, p. 168–170, 2 maio 2002.

KHAMIS, A.; RAOULT, D.; SCOLA, B. LA. rpoB Gene Sequencing for Identification of Corynebacterium Species rpoB Gene Sequencing for Identification of Corynebacterium Species. v. 42, n. 9, 2004.

KHAMIS, A.; RAOULT, D.; SCOLA, B. LA. Comparison between rpoB and 16S rRNA Gene Sequencing for Molecular Identification of 168 Clinical Isolates of Corynebacterium Comparison between rpoB and 16S rRNA Gene Sequencing for Molecular Identification of 168 Clinical Isolates of Corynebacterium. 2005.

LARTIGUE, M.; MONNET, X.; FLÈCHE, A. LE; *et al.* Corynebacterium ulcerans in an immunocompromised patient with diphtheria and her dog. **Journal of clinical …**, v. 43, n. 2, p. 999–1001, 2005.

RIEGEL, P.; RUIMY, R.; BRIE, D. DE; *et al.* Taxonomy of Corynebacterium diphtheriae and related taxa , with recognition of Corynebacterium ulcerans sp . nov . nom . rev . v. 126, p. 271–276, 1995.

SEKIZUKA, T.; YAMAMOTO, A.; KOMIYA, T.; *et al.* Corynebacterium ulcerans 0102 carries the gene encoding diphtheria toxin on a prophage different from the C. diphtheriae NCTC 13129 prophage. **BMC microbiology**, v. 12, p. 72, jan 2012.

SILVA, A. S. S.; BARAÚNA, R. A.; SÁ, P. C. G.; *et al.* Draft Genome Sequence of Corynebacterium ulcerans FRC58 , Isolated from the Bronchitic Aspiration of a Patient in France. v. 2, n. 1, p. 1–2, 2014.

SING, A.; BIERSCHENK, S.; HEESEMANN, J. Classical diphtheria caused by Corynebacterium ulcerans in Germany: amino acid sequence differences between diphtheria toxins from Coryebacterium diphteriae and C. ulcerans. **Clinical infectious diseases : an official publication of the Infectious Diseases Society of America**, v. 40, p. 325–326, 2004.

SOARES, S. C.; SILVA, A.; TROST, E.; *et al.* The Pan-Genome of the Animal Pathogen Corynebacterium pseudotuberculosis Reveals Differences in Genome Plasticity between the Biovar ovis and equi Strains. **PloS one**, v. 8, n. 1, p. e53818, jan 2013.

TROST, E.; AL-DILAIMI, A.; PAPAVASILIOU, P.; *et al.* Comparative analysis of two complete Corynebacterium ulcerans genomes and detection of candidate virulence factors. **BMC genomics**, v. 12, n. 1, p. 383, jan 2011.

TROST, E.; BLOM, J.; SOARES, S. D. C.; *et al.* Pangenomic study of Corynebacterium diphtheriae that provides insights into the genomic diversity of pathogenic isolates from cases of classical diphtheria, endocarditis, and pneumonia. **Journal of bacteriology**, v. 194, n. 12, p. 3199–215, jun 2012.

VIANA, M. V. C.; BENEVIDES, L. J.; MARIANO, D. C. B.; *et al.* Genome Sequence of Corynebacterium ulcerans Strain 210932. v. 2, n. 6, p. 6–7, 2014.

WAGNER, K. S.; WHITE, J. M.; LUCENKO, I.; *et al.* Diphtheria in the Postepidemic. v. 18, n. 2, p. 217–225, 2012.

WELLINGHAUSEN, N.; SING, A.; KERN, W. V; *et al.* Case report A fatal case of necrotizing sinusitis due to toxigenic Corynebacterium ulcerans. v. 63, p. 59–63, 2002.

ZOYSA, A. DE; HAWKEY, P. M.; ENGLER, K.; *et al.* Characterization of toxigenic Corynebacterium ulcerans strains isolated from humans and domestic cats in the United Kingdom. **Journal of clinical microbiology**, v. 43, n. 9, p. 4377–81, set 2005.

VIII. Appendices

VIII.1. Curriculum Vitae

Address to this CV	http://lattes.cnpq.br/1991446268436258
Full name	Luis Carlos Guimarães
Name used in Bibliographic Citations	GUIMARÃES, L. C.; Guimaraes, L. C.; Guimarães, Luis C.; Guimarães, Luis; Luis Carlos Guimarães; Carlos Guimarães L; GUIMARÃES, LUÍS C.; GUIMARÃES, LUIS CARLOS; Carlos Guimaraes, Luis; CARLOS GUIMARÃES, LUIS; GUIMARÃES, LUÍS CARLOS; GUIMARÃES, LUIS C; GUIMARÃES, L.C.
Parental information	Geraldo Ferreira Guimarães and Nilva Maria do Amaral Guimarães
Birth information	31/05/1983 - Patos de Minas/MG - Brazil
Identification document	MG11062949 SSP - MG - 25/11/1996
CPF Number	059.608.546-00
Passport	FE375437
Professional Address	Universidade Federal de Minas Gerais Instituto de Ciências Biológicas Departamento de Biologia Geral Av. Antônio Carlos, ICB bloco Q3 259 31270-215, MG - Brazil
Phone number	+55 31 3409-2610
Eletronic Address	luisguimaraes.bio@gmail.com

Formal Education

2011 – current date
Doctorate in Genetics.
Universidade Federal de Minas Gerais, UFMG, Belo Horizonte, Brazil with Sandwich Doctorate in Bielefeld University (Advisor : Andreas Tauch)
Title: Comparative Genomics and Pan-Genomic Study of genus *Corynebacterium*
Advisor: Vasco Ariston de Carvalho Azevedo
Co-orientador: Siomar de Castro Soares

2008 - 2011
Master's in Genetics (emphasis in Bioinformatics)
Universidade Federal do Pará, UFPA, Belem, Brazil
Title: Modelagem por homologia do proteoma predito de Corynebacterium pseudotuberculosis e dinâmica molecular da Metaloendopeptidase deste organismo, Year of degree: 2011
Advisor: Vasco Ariston de Carvalho Azevedo

2005 - 2007
Bachelor's in Biology.
Centro Universitário de Patos de Minas, UNIPAM, Patos De Minas,

Brazil
Title: Avaliação do potencial genotóxico ou antigenotóxico do ômega-3 nas células somáticas das asas da Drosophila melanogaster
Advisor: Júlio César Nepomuceno

Areas of Expertise 1. Molecular Genetics of Microorganisms
2. Genomics
3. Bioinformatics

Bibliographic Production

Articles Published in Scientific Journals

1. Barauna, R. A., **Guimaraes, L. C.**, Veras, A. A. O., de Sa, P. H. C. G., Gracas, D. A., Pinheiro, K. C., Silva, A. S. S., Folador, E. L., Benevides, L. J., Viana, M. V. C., Carneiro, A. R., Schneider, M. P. C., Spier, S. J., Edman, J. M., Ramos, R. T. J., Azevedo, V., Silva, A. Genome Sequence of *Corynebacterium pseudotuberculosis* MB20 bv. *equi* Isolated from a Pectoral Abscess of an Oldenburg Horse in California. Genome Announcements. , v.2, p.e00977-14 - e00977-14, 2014.

2. Oliveira, L. C., Saraiva, T. D. L., Soares, S. C., Ramos, R. T. J., Sa, P. H. C. G., Carneiro, A. R., Miranda, F., Freire, M., Renan, W., Junior, A. F. O., Santos, A., Pinto, A. C., Souza, B. M., Castro, C. P., Diniz, C. A. A., Rocha, C. S., Mariano, D. C. B., Aguiar, E. L., Folador, E. L., Barbosa, E. G. V., Aburjaile, F. F., Gonçalves, L. A., **Guimarães, L. C.**, Azevedo, M., Agresti, P. C. M., Silva, R. F., Tiwari, S., Almeida, S. S., Hassan, S. S., Pereira, V. B., Abreu, V. A. C., Pereira, U. P., Dorella, F. A., Carvalho, A. F., Pereira, F. L., Leal, C. A. G., Figueiredo, H. C. P., Silva, A., Miyoshi, A., Azevedo, V. Genome Sequence of *Lactococcus lactis* subsp. *lactis* NCDO 2118, a GABA-Producing Strain. Genome Announcements. , v.2, p.e00980-14 - e00980-14, 2014.

3. Hassan, S. S., Tiwari, S., **Guimarães, L. C.**, Jamal, S. B., Folador, E. L., Sharma, N. B., Soares, S. C., Almeida, S. S., Ali, A., Povoa, F. D., Abreu, V. A. C., Jain, N., Bhattacharya, A., Juneja, L., Miyoshi, A., Silva, A., Barh, D., Turjanski, A. G., Azevedo, V., Ferreira, R. S. Proteome scale comparative modeling for conserved drug and vaccine targets identification in *Corynebacterium pseudotuberculosis*. BMC Genomics. Fator de Impacto (2013 JCR): 4,0410, v.15, p.S3 - , 2014.

4. **Guimaraes, L. C.**, Soares, S. C., Albersmeier, A., Blom, J., Jaenicke, S., Azevedo, V., Soriano, F., Tauch, A., Trost, E. Complete Genome Sequence of *Corynebacterium urealyticum* Strain DSM 7111, Isolated from a 9-Year-Old Patient with Alkaline-Encrusted Cystitis. Genome Announcements. , v.1, p.e00264-13 - e00264-13, 2013.

5. Pereira, U P., Santos, A. R., Hassan, S S., Aburjaile, F. F., Soares, S. C., Ramos, R. T. J., Carneiro, A. R., **Guimarães, L. C.**, Almeida,S. S., Diniz, C. A. A., Barbosa, M. S., Sá, P. G., Ali, A., Bakhtiar, S. M., Dorella, F. A., Zerlotini, A., Araújo, F. M. G., Leite, L. R., Oliveira, G., Miyoshi, A., Silva, A., Azevedo, V., Figueiredo, H. C. P. Complete genome sequence of *Streptococcus agalactiae* strain SA20-06, a fish pathogen associated to meningoencephalitis outbreaks. Standards in Genomic Sciences. Fator de Impacto(2013 JCR): 3,1670, v.8, p.188 - 197, 2013.

6. Barh, D., Gupta, K., Jain, N., Khatri, G., León-Sicairos, N., Canizalez-Roman, A., Tiwari, S., Verma, A., Rahangdale, S., Shah Hassan, S., Santos, A. R., Ali, A., **Guimarães, L. C.**, Ramos, R. T. J., Devarapalli, P., Barve, N., Bakhtiar, M., Kumavath, R., Ghosh, P.,

Miyoshi, A., Silva, A., Kumar, A, Misra, A. N., Blum, K., Baumbach, J., Azevedo, V. Conserved host-pathogen PPIs : Globally conserved inter-species bacterial PPIs based conserved host-pathogen interactome derived novel target in *C. pseudotuberculosis, C. diphtheriae, M. tuberculosis, C. ulcerans, Y. pestis*, and *E. coli* targeted by Piper betel compounds. Integrative Biology. Fator de Impacto(2013 JCR): 3,9960, v.5, p.495 - , 2013.

7. Pereira, U.P., Soares, S.C., Blom, J., Leal, C.A.G., Ramos, R.T.J., **Guimarães, L.C.**, Oliveira, L.C., Almeida, S.S., Hassan, S.S., Santos, A.R., Miyoshi, A., Silva, A., Tauch, A., Barh, D., Azevedo, V., Figueiredo, H.C.P. In silico prediction of conserved vaccine targets in *Streptococcus agalactiae* strains isolated from fish, cattle, and human samples. Genetics and Molecular Research. Fator de Impacto(2013 JCR): 0,8500, v.12, p.2902 - 2912, 2013.

8. Dorella, F. A., Gala-Garcia, A., Pinto, A. C., Sarrouh, B., Antunes, C. A., Ribeiro, D., Aburjaile, F., Fiaux, K. K., **Guimarães, L. C**, Seyffert, N., El-Aouar, R. A., Silva, R., Hassan, S. S., Castro, T. L. P., Marques, W. S., Ramos, R., Carneiro, A., Sá, P., Miyoshi, A., Azevedo, V., Silva, A. Progression of 'OMICS' methodologies for understanding the pathogenicity of *Corynebacterium pseudotuberculosis*: the Brazilian experience. Computational and Structural Biotechnology Journal. , v.6, p.1 - 7, 2013.

9. Soares, S. C., Silva, A, Trost, E., Blom, J., Ramos, R., Carneiro, A., Ali, A., Santos, A. R., Pinto, A. C., Diniz, C., Barbosa, E. G. V., Dorella, F. A., Aburjaile, F., Rocha, F. S., Nascimento, K. K. F., **Guimarães, L. C.**, Almeida, S., Hassan, S. S., Bakhtiar, S. M., Pereira, U. P., Abreu, V. A. C., Schneider, M. P. C., Miyoshi, A., Tauch, A., Azevedo, V. The Pan-Genome of the Animal Pathogen *Corynebacterium pseudotuberculosis* Reveals Differences in Genome Plasticity between the Biovar *ovis* and *equi* Strains. Plos One. Fator de Impacto(2013 JCR): 3,5340, v.8, p.e53818 - , 2013.

10. Ali, A., Soares, S. C., Santos, A. R., **Guimarães, L. C.**, Barbosa, E., Almeida, S. S., Abreu, V. A. C., Carneiro, A. R., Ramos, R. T. J., Bakhtiar, S. M., Hassan, S S., Ussery, D W., On, S., Silva, A., Schneider, M. P., Lage, A. P., Miyoshi, A., Azevedo, V. *Campylobacter fetus* subspecies: Comparative genomics and prediction of potential virulence targets. Gene (Amsterdam). Fator de Impacto(2013 JCR): 2,0820, v.508, p.145 - 156, 2012.

11. Hassan, S. S., **Guimarães, L. C.**, Pereira, U. P., Islam, A., Ali, A., Bakhtiar, S. M., Ribeiro, D., Santos, A. R., Soares, S. C., Dorella, F. A., Pinto, A. C., Schneider, M. P. C., Barbosa, M. S., Almeida, S., Abreu, V., Aburjaile, F., Carneiro, A. R., Cerdeira, L. T., Fiaux, K., Barbosa, E., Diniz, C., Rocha, F. S., Ramos, R. T. J., Jain, N., Tiwari, S., Barh, D., Miyoshi, A., MÜLLER, B., Silva, A., Azevedo, V. Complete genome sequence of *Corynebacterium pseudotuberculosis* biovar *ovis* strain P54B96 isolated from antelope in South Africa obtained by rapid next generation sequencing technology. STAND GENOMIC SCI. Fator de Impacto(2013 JCR): 3,1670, v.7, p.189 - 199, 2012.

12. Silva, A., Ramos, R. T. J., Carneiro, A. R., Pinto, A. C., Soares, S. C., Santos, A. R., Almeida, S. S., **Guimaraes, L. C.**, Aburjaile, F. F., Barbosa, E. G. V., Dorella, F. A., Rocha, F. S., Lopes, T. S., Kawasaki, R., Sá, P. G., Coimbra, N. A. R., Cerdeira, L. T., Barbosa, M. S., Schneider, M. P. C., Miyoshi, A., Selim, S. A. K., Moawad, M. S., Azevedo, V. Complete Genome Sequence of *Corynebacterium pseudotuberculosis* Cp31, Isolated from an Egyptian Buffalo. Journal of Bacteriology (Print). Fator de Impacto(2013 JCR): 2,6880, v.194, p.6663 - 6664, 2012.

13. Silva, A., Lopes, T., Ramos, R. T. J., Carneiro, A. R., Dorella, F. A., Rocha, F. S., Santos, A. R., **Guimarães, L C.**, Barbosa, E. G. V., Ribeiro, D., Fiaux, K., Diniz, C. A. A., Abreu, V. A. C., Almeida, S. S., Hassan, S., Amjad A., Bakhtiar, S., Aburjaile, F. F., Pinto, A. C., Soares, S. C., Pereira, U , Schneider, M. P. C., Miyoshi, A., Edman, J , Azevedo, V.

Complete Genome Sequence of *Corynebacterium pseudotuberculosis* Strain Cp267, Isolated from a Llama. Journal of Bacteriology (Print). Fator de Impacto(2013 JCR): 2,6880, v.194, p.3567 - 3568, 2012.

14. Pethick, F E, Lainson, A F, Yaga, R, Flockhart, A, Smith, D G E, Donachie, W, Cerdeira, L T, Silva, A , Bol, E, Lopes, T S, Barbosa, M S, Pinto, A C, Santos, A R, Soares, S. C., Almeida, S. S., **Guimaraes, L. C.**, Aburjaile, F. F., Abreu, V. A. C., Ribeiro, D., Fiaux, K.K., Diniz, C. A. A., Barbosa, E. G. V., Pereira, U. P., Hassan, S. S., Ali, A., Bakhtiar, S. M., Dorella, F. A., Carneiro, A. R., Ramos, R. T. J., Rocha, F. S., Schneider, M. P. C., Miyoshi, A., Azevedo, V. and Fontaine, M. C. Complete Genome Sequence of *Corynebacterium pseudotuberculosis* Strain 1/06-A, Isolated from a Horse in North America. Journal of Bacteriology (Print). Fator de Impacto(2013 JCR): 2,6880, v.194, p.4476 - 4476, 2012.

15. Pethick, F E, Lainson, A F, Yaga, R, Flockhart, A, Smith, D G E, Donachie, W, Cerdeira, L T, Silva, A, Bol, E, Lopes, T S, Barbosa, M S, Pinto, A C, Santos, A R, Soares, S. C., Almeida, S. S., **Guimaraes, L. C.**, Aburjaile, F. F., Abreu, V. A. C., Ribeiro, D. Fiaux, K. K., Diniz, C. A. A., Barbosa, E. G. V., Pereira, U. P., Hassan, S. S., Ali, A., Bakhtiar, S. M., Dorella, F. A., Carneiro, A. R., Ramos, R. T. J., Rocha, F. S., Schneider, M. P. C., Miyoshi, A., Azevedo, V. and Fontaine, M. C. Complete Genome Sequences of *Corynebacterium pseudotuberculosis* Strains 3/99-5 and 42/02-A, Isolated from Sheep in Scotland and Australia, Respectively. Journal of Bacteriology (Print). Fator de Impacto(2013 JCR): 2,6880, v.194, p.4736 - 4737, 2012.

16. Soares, S. C., Trost, E., Ramos, R. T. J., Carneiro, A. R., Santos, A. R., Pinto, A. C., Barbosa, E., Aburjaile, F., Ali, A., Diniz, C. A. A., Hassan, S. S., Fiaux, K., **Guimarães, L. C.**, Bakhtiar, S. M., Pereira, U., Almeida, S. S., Abreu, V. A. C., Rocha, F. S., Dorella, F. A., Miyoshi, A., Silva, A., Azevedo, V., Tauch, A. Genome sequence of *Corynebacterium pseudotuberculosis* biovar *equi* strain 258 and prediction of antigenic targets to improve biotechnological vaccine production. Journal of Biotechnology. Fator de Impacto(2013 JCR): 2,8840, v.164, p.5 - , 2012.

17. Carneiro, A R, Ramos, R T J, Dall'Agnol, H, Pinto, A C, Soares, S C, Santos, A R, **Guimaraes, L C**, Almeida, S S, Barauna, R A, Gracas, D A et al Genome Sequence of *Exiguobacterium antarcticum* B7, Isolated from a Biofilm in Ginger Lake, King George Island, Antarctica. Journal of Bacteriology (Print). Fator de Impacto(2013 JCR): 2,6880, v.194, p.6689 - 6690, 2012.

18. Ramos, R T J, Silva, A, Carneiro, A R, Pinto, A C, Soares, S C, Santos, A R, Almeida, S S, **Guimaraes, L C**, et al Genome Sequence of the *Corynebacterium pseudotuberculosis* Cp316 Strain, Isolated from the Abscess of a Californian Horse. Journal of Bacteriology (Print). Fator de Impacto(2013 JCR): 2,6880, v.194, p.6620 - 6621, 2012.

19. D'Afonseca, V., Soares, S. C., Ali, A., Santos, A. R., Pinto, A. C., Magalhães, A. A. C., Faria, C. J., Barbosa, E., **Guimarães, L. C.**, Eslabão, M., Almeida, S. S., Abreu, V. A. C., Zerlotini, A., Carneiro, A. R., Cerdeira, L. T., Ramos, R. T. J., Hirata Jr, R., Mattos-Guaraldi, A. L., Trost, E. Tauch, A. Silva, A., Schneider, M. P. C., Miyoshi, A., Azevedo, V. Reannotation of the *Corynebacterium diphtheriae* NCTC13129 genome as a new approach to studying gene targets connected to virulence and pathogenicity in diphtheria. Open Access Bioinformatics. , v.2012:4, p.1 - 13, 2012.

20. **Guimarães, L. C.**, Silva, N. F., Miyoshi, A., Schneider, M. P.C., Silva, A., Azevedo, V., Brasil, D. S. B., Lameira, J., Alves, C. N. Structure modeling of a metalloendopeptidase from *Corynebacterium pseudotuberculosis*. Computers in Biology and Medicine. Fator de Impacto(2013 JCR): 1,4750, v.193, p.S00104825120001 - , 2012.

21. Santos, A. R., Carneiro, A., Gala-García, A., Pinto, A., Barh, D., Barbosa, E., Aburjaile, F., Dorella, F. Rocha, F., **Guimarães, L.**, Zurita-Turk, M., Ramos, R., Almeida, S., Soares, S., Pereira, U., Abreu, V. C., Silva, A., Miyoshi, A. and Azevedo, V. The *Corynebacterium pseudotuberculosis* in silico predicted pan-exoproteome. BMC Genomics. Fator de Impacto(2013 JCR): 4,0410, v.13, p.S6 - , 2012.

22. Ramos, R. T. J., Carneiro, A. R., Soares, S. C., Santos, A. R., Almeida, S., **Guimarães, L.**, Figueira, F., Barbosa, E., Tauch, A., Azevedo, V., Silva, A. Tips and tricks for the assembly of a *Corynebacterium pseudotuberculosis* genome using a semiconductor sequencer. Microbial Biotechnology (Online). , v.n/a, p.n/a - n/a, 2012.

23. Hassan, S S., Schneider, M P C, Ramos, R T J, Carneiro, A R, Ranieri, A, **Guimaraes, L C**, Ali, A , Bakhtiar, S M, Pereira, U P, Santos, A R et al Whole-Genome Sequence of *Corynebacterium pseudotuberculosis* Strain Cp162, Isolated from Camel. Journal of Bacteriology (Print). Fator de Impacto(2013 JCR): 2,6880, v.194, p.5718 - 5719, 2012.

24. Barh, D., Jain, N., Tiwari, S., Li, L., D'Afonseca, V., Ali, A., Santos, A. R., **Guimarães, L. C.**, Soares, S. c., Miyoshi, A., Bhattacharjee, A., Misra, A. N., Silva, A., Kumar, A., Azevedo, V. A novel comparative genomics analysis for common drug and vaccine targets in *Corynebacterium pseudotuberculosis* and other CMN group of human pathogens. Chemical Biology & Drug Design (Print). Fator de Impacto(2013 JCR): 2,5070, v.78, p.73 - 84, 2011.

25. Cerdeira, L. T., Schneider, M. P. C., Pinto, A. C., Almeida, S. S., Santos, A. R., Barbosa, E. G. V., Ali, A., Aburjaile, F. F., V. A. C. Abreu, V. A. C., **Guimarães, L. C.**, Soares, S. C., Dorella, F. A., Rocha, F. S., Bol, E., Sá, P. H. C. G., Lopes, T. S., Barbosa, M. S., Carneiro, A. R., Ramos, R. T. J., Coimbra, N. A. R., Lima, A. R. J., Barh, D., Jain, N., Tiwari, S., Raja, R. Zambare, V., Ghosh, P., Trost, E., Tauch, A., Miyoshi, A., Azevedo, V. and Silva, A. Complete Genome Sequence of *Corynebacterium pseudotuberculosis* Strain CIP 52.97, Isolated from a Horse in Kenya. Journal of Bacteriology (Print). Fator de Impacto(2013 JCR): 2,6880, v.193, p.7025 - 7026, 2011.

26. Cerdeira, L. T., Pinto, A. C., Schneider, M. P. C., Almeida, S.S., Santos, A. R., Barbosa, E. G. V., Ali, A., Barbosa, M. S., Carneiro, A. R., Ramos, R. T. J., Oliveira, R. S., Barh, D., Barve, N., Zambare, V., Belchior, S. E., **Guimarães, L. C.**, Soares, S. C., Dorella, F. A., Rocha, F. S., Abreu, V. A. C., Tauch, A. Trost, A., Miyoshi, A., Azevedo, V. and Silva, A. Whole-Genome Sequence of *Corynebacterium pseudotuberculosis* PAT10 Strain Isolated from Sheep in Patagonia, Argentina. Journal of Bacteriology (Print). Fator de Impacto(2013 JCR): 2,6880, v.193, p.6420 - 6421, 2011.

Chapters published

Silva, A.; Ramos, R. T. J.; Carneiro, A. R.; Almeida, S. S.; Abreu, V. A. C.; Santos, A. R.; Soares, S. C.; Pinto, A. C.; **Guimaraes, L. C.**; Barbosa, E. G. V.; Schneider, M. P. C.; Zambare, V. ; Barh, D.; Miyoshi, A.; Azevedo, V. Next-Generation Sequencing and Assembly of Bacterial Genomes. In: Debmalya Barh; Vasudeo Zambare; Vasco Azevedo. (Org.). OMICS: Applications in Biomedical, Agricultural, and Environmental Sciences. 1ed.Inglaterra: CRC Press, 2013, v. , p. 1-713.

Articles in Magazines

Soares, SC; Silva, A; Ramos, RTJ; Cerdeira, L; Ali, A; Santos, AR; Pinto, AC; Cassiano, AAM; Aburjaile, FF; Carneiro, AR; **Guimarães, LC**; Barbosa, EGV; Almeida, SS; Abreu, VAC; Miyoshi, A; Azevedo, V. Plasticidade Genômica e Evolução Bacteriana. Microbiologia in Foco, 26 º CBM - Foz do Iguaçu, v. 16, p. 31-8, 2011.