

**Comparative Genomics of *Bacillus* spp.: can core genome be the key for new therapeutics?  
Assessment of the core genome of *Bacillus* spp. And its stability**

Carlos Alonso<sup>1,2</sup>, Francisco M. Couto<sup>3</sup> & Ricardo Dias<sup>1</sup>

<sup>1</sup>LMB-BioFIG Laboratory for Microbiology and Biotechnology – Center for Biodiversity, Functional & Integrative Genomics

<sup>2</sup>Laboratório de Bromatologia e Defesa Biológica, Exército Português

<sup>3</sup>LASIGE, Department of Informatics, Faculty of Sciences, University of Lisbon

Anthrax is a disease caused by the bacteria *Bacillus anthracis* which primarily infects ruminants, but may occasionally infect human beings. This disease has been nearly eradicated in western countries in the last decades, but has been brought back to light following its use in the terrorist attacks in 2001 in the USA. Given this bacteria's ability to survive in the environment and its easy dissemination thanks to its spore forming ability, it is an ideal agent to be used in a bioterrorist attack. In addition to the bacteria's innate pathogenic characteristics, taking in account the current state of molecular biology, there is a chance that a *B. anthracis* strain used in a bioterrorist attack has been genetically modified to gain antimicrobial resistance mechanisms. With the advent of synthetic biology there is also the risk of *B. anthracis* pathogenic characteristics being introduced in a different *Bacillus* species. This would create a new pathogen with the same pathogenic characteristics as *B. anthracis*, but which would not be identified by the detection methods for *B. anthracis* and might need a different clinical treatment. The objective of this work is to identify the core genome and essential genes of the genus *Bacillus*. The present work aims in its first phase to identify the core genome of the genus *Bacillus*. Within the scope of this work, the core genome is considered as the genetic information which is transversally present in most or all strains of the same genus. The fact that this genetic material remains conserved in divergent species suggests that it is essential for survival. In order to identify the core genome of the genus *Bacillus*, all the completely sequenced genomes of the genus *Bacillus* available in NCBI will be used, along with a set of controls, and a clustering algorithm will be used to identify sets of orthologous genes conserved in all or most strains of the genus *Bacillus*. The second phase of the project will consist in the identification of essential genes in *Bacillus* spp. Essential genes are genes that are indispensable to support cellular life. These genes constitute a minimal gene set required for a living cell. The identification of essential genes will be made through comparison with known essential genes, the assessment of several parameters which may reflect essentiality – such as function, Ka/Ks ratio, gene localization, proteic and aminoacidic parameters – and machine learning methods. The knowledge of the set of core and essential genes in the *Bacillus* genus would make way for the subsequent task of selecting a subset of target genes, which would be used to develop molecules with therapeutic or prophylactic action in *B. anthracis*. Seeing as these molecules would interact with essential, conserved proteins, it would be more difficult for someone to engineer a resistant strain, or for the bacteria to naturally develop resistance mechanisms. In a preliminary run with 80 proteomes (57 *Bacillus* spp. Proteomes and 23 from other bacterial genus as control), 289405 proteins, it was observed that 210153 proteins grouped in 27272 clusters, from which 20791 were found in the *Bacillus* genus. 18468 proteins were assigned to clusters (n = 324) including all *Bacillus* proteomes analyzed. Analysis of the main functional classes will be presented.