TITLE: STRUCTURAL DIVERSITY OF BACTERIAL PAN-GENOMES

AUTHORS: COSTA, S.S.[1]; GUIMARÃES, L.C.[1]; SOARES, S.C.[2]; SILVA, A.[1,3]; BARAÚNA, R.A.[1,3]

INSTITUTION : [1] CENTRO DE GENÔMICA E BIOLOGIA DE SISTEMAS, UNIVERSIDADE FEDERAL DO PARÁ, 66075-110, BELÉM-PA, BRAZIL.
[2] INSTITUTO DE CIÊNCIAS BIOLÓGICAS E NATURAIS, UNIVERSIDADE FEDERAL DO TRIÂNGULO MINEIRO., RUA GETÚLIO GUARITÁ, S/N, UBERABA-MG, BRASIL. 38025-180.
[3] LABORATÓRIO DE ENGENHARIA BIOLÓGICA, PARQUE DE CIÊNCIA E TECNOLOGIA GUAMÁ, 66075-750, BELÉM, BRAZIL.

ABSTRACT:
Bacteria are cosmopolitan organisms. Pan-genome studies provide insights into how the gene content is related to the lifestyle and even bacterial pathogenicity. This work aimed to analyze the pan-genome composition of nine bacterial species with the higher number of complete genomes deposited in GenBank. The analyzed species were: *Escherichia coli*, *Bordetella pertussis*, *Campylobacter jejuni*, *Streptococcus pneumonia*, *Staphylococcus aureus*, *Klebsiella pneumonia*, *Listeria monocytogenes*, *Mycobacterium tuberculosis*, and *Pseudomonas aeruginosa*. We also analyzed the influence of uncharacterized proteins on the pan-genome structure. Due to the difficulty in determining the function of several uncharacterized proteins it is reasonable to affirm that some of these proteins are actually a false-positive prediction of a gene and therefore may have great influence on the pan-genome calculation. Fifty strains of each species were randomly selected and downloaded from the GenBank database totalizing 450 genomes. The gene annotation was normalized submitting all genomes to the RAST server. The pan-genome was calculated in PGAP using the parameters of 50% identity, 60% coverage, and e-value 10e-5. The same analysis was performed after excluding the uncharacterized proteins using Artemis genome browser. Removal of uncharacterized proteins altered the alpha value of Heap's Law in all species analyzed. Thus, these proteins have a strong influence on the genome plasticity. Nevertheless, all species remained with an open pan-genome ($\alpha<1$). As expected, this removal also affected the gene density of all genomes decreasing from 1,000 genes/Mbp to 667 genes/Mbp. About 26% of genes from *Staphylococcus aureus* were uncharacterized proteins. The lowest number was observed in *Escherichia coli* where about 13% of genes were uncharacterized proteins. *Klebsiella pneumonia* and *Escherichia coli* showed the strongest modification on their pan-genome structure after the removal of uncharacterized proteins, resulting in a reduction of 36% and 29% of the pan-genome size, respectively. *Bordetella pertussis* and *Listeria monocytogenes* were the species with an alpha value closer to 1. This work showed a large-scale analysis of the pan-genome structure of the nine most representative species in databases. We were able to demonstrate that uncharacterized proteins directly affect the structure of the pan-genome with different impacts depending on the bacterial species.

Keywords: Pan-genome, genomics, comparative genomics, heap's law.

Development Agency: CAPES – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior; FAPESPA – Fundação Amazônia de Amparo a Estudos e Pesquisa.