TITLE: Integrating Overall Genome-Relatedness Indexes (OGRIs) to Improve Genome-Based Species Identification of Bacterial Pathogens

AUTHORS: ALVES, D. A., AGUIAR, E. R. G. R, PACHECO, L. G. C.

INSTITUITION: Universidade Federal da Bahia, Salvador, Bahia (Avenida Reitor Miguel Calmon, s/n. Vale do Canela 40110-100 – Salvador - BA, BRASIL); Universidade Estadual de Santa Cruz, Ilhéus, Bahia (Campus Soane Nazaré de Andrade, Rod. Jorge Amado, Km 16, Salobrinho, Ilhéus – Ba, BRASIL)

ABSTRACT

BACKGROUND: Overall Genome-Relatedness Indexes (OGRIs) have been extensively used in recent years for taxonomic classification of bacteria. Average Nucleotide Identity by BLAST (ANIb) in pairwise genome comparisons is widely regarded as the most accurate index for bacterial species circumscription, when considering a species boundary of ca. 95-96% identity. However, the sole use of ANIb for species identification may render confusing results for some closely related bacterial groups, such as many pathogenic *Corynebacterium* spp. AIM: To evaluate the performance of an in-house developed species-classifier, based on the correlation values between different genome relatedness indexes, to correctly classify genomic sequences from the *Corynebacterium diphtheriae* group. METHODS: 213 genomic sequences corresponding to three closely related *Corynebacterium* species were retrieved from NCBI's Genome DB: 188 classified as *C. diphtheriae* and 03 isolates of *C. diphtheriae* subsp. Lausannense; 10 as *Corynebacterium belfantii*; 01 as *Corynebacterium rouxii* and 11 classified as *C. diphtheriae* obtained from European Nucleotide Archive. Tetranucleotide usage patterns (TETRA) and ANIb were calculated through the JSpecies Web server application and compared all-vs-all. Resulting matrices were merged to generate a single fingerprint matrix, which was used to compute Spearman's correlation values among bacterial genomes using an in-house script developed on R software. Multilocus sequence analysis (MLSA) with six house-keeping genes and split-decomposition analyses were used to confirm relationships between the various species. RESULTS: In total, 41,209 genome-to-genome comparisons composed the fingerprint matrix that was used to build a dendrogram with well-defined clades (>95% bootstrap confidence). The groups containing *C. belfantii* and *C. rouxii* were clearly distinguished by this strategy, as opposed to the use of ANIb alone. Additionally, we observed that our results are corroborated with the MLSA, highlighting the mistaken classification in the NCBI. We also observed that Lausannense subspecies form demarcated clades of diphtheria. CONCLUSIONS: An in-house developed classifier that integrates different OGRIs results was the most efficient tool for species circumscription in the *C. diphtheriae* group, when compared to ANIb alone. We anticipate that this new strategy can be extrapolated to improve genome-based identification of other clinically important bacterial pathogens.

E-mail: Daniele Almeida Alves: almeidaniele27@gmail.com

Keywords: Emerging pathogens, Genomic taxonomy, Average Nucleotide Identity, Tetranucleotide patterns, *Corynebacterium* spp.

Development Agency: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior