TITLE: IDENTIFICATION OF NOVEL GENE PRODUCTS IN *Bifidobacterium breve* GENOME USING NGS DATA

AUTHORS: DAMASCENO, B. A.[1]; OLIVEIRA, M. S. [2]; DIAS, L.M.[3]; VERAS, A. A. O.[2]; SÁ, P. H. C. G.[4]; ALVES, J. T. C.[1]

INSTITUTION: [1]PARÁ STATE UNIVERSITY (AV. HILÉIA, S/N, CEP 68502-100, MARABÁ, PARÁ, BRAZIL); [2] FEDERAL UNIVERSITY OF PARÁ (RUA ITAIPU, 36, CEP 68464-000, TUCURUÍ, PARÁ, BRAZIL); [3] FEDERAL UNIVERSITY OF PARÁ (AUGUSTO CORRÊA, 01, CEP 66075-110, BELEM, PARÁ, BRAZIL); [4]FEDERAL RURAL UNIVERSITY OF AMAZONIA (RODOVIA PA 140, CEP 68680-000, TOMÉ-AÇU, PARÁ, BRAZIL);

ABSTRACT:

The representation of the genes of an organism *in silico* is essential for bioinformatics analyzes since this information will serve as the basis for several omic analyzes. One of the first post-sequencing analyzes is the *de novo* genome assembly, in which it represents the genome of the organism sequenced from the reads. During the assembly the genomic content is often not fully represented due to errors inherent in the assembly process or problems related to the coverage of the data. Consequently, they influence the identification of some genes in the genome, both in complete genomes and in drafts genomes, thus impairing gene annotation analyzes. *Bifidobacterium breve* is Gram positive bacteria, commonly encountered in the gastrointestinal tract of mammals, including humans, and acts as a beneficial probiotic. Thus the aim of this work is to identify the new gene products not represented in the genome of *Bifidobacterium breve* using NGS data. The fasta reference genome used was *B. breve* DSM 20213 (acession number: NZ_AP012324.1) available at NCBI. The reads used are from the complete genome sequencing on the Illumina MiSeq platform, available on EMBL-EBI (acession number: SRR5310870). The ImproveAssembly computational tool was used to identify the new products. This tool analyzes the reads that have not been mapped against the reference genome, performing a reassembly of that data and a later annotation. Subsequently, the identified gene products are analyzed by Blast against the products of the reference genome to identify new products not represented in the original genome. As the result, a total of 31 new gene products were identified that are not represented in the reference genome of *B. breve* DSM 20213 in which 12 products are hypothetical proteins and 19 are products with function described, among them “NADH peroxidase Npx”, “Similar to tetracycline resistance protein” and “Beta-lactamase class C and other penicillin binding proteins” the latter is very important because it is a gene that gives bacteria resistance to antibiotics. Thus, in addition to several new gene products identified, it was also possible to validate that even in complete genomes the total gene content may not be fully represented, which may influence to subsequent analyzes.

Keywords: *Bifidobacterium breve,* bioinformatics, NGS, novel gene products.