Evolution of the endonuclease *nth* in Lactobacillales species

Integrated 1st Cycle Project in Biological Engineering

Beatriz Labiza, 87441

Universidade de Lisboa, Instituto Superior Técnico, Lisboa

BSc Biological Engineering

Introduction

Lactic acid bacteria (LAB) are an heterogeneous group of microorganisms that convert carbohydrates into lactic acid (Alexander Bolotin, 2001). Taxonomically, LAB are comprised in the Lactobacillales order (Annex 1), which is divided into six taxonomic families: the Carnobactereaceae, Streptococcaceae, Enterococcaceae, Leuconostocaceae, Lactobacillaceae and Aerococacceae. The Streptococcaceae family comprises two distinct genera, the *Streptococcus* and *Lactococcus* (Annex 2).

In 2001, the genome of the *Lactococcus lactis* IL1403 was sequenced, opening the door to the use of genome-wide Omics approaches to characterize this important microbial factory workhorse. (Alexander Bolotin, 2001)

Lactococcus lactis has gained interest in research and in the industry. On the other hand, the non-pathogenic LAB are a group of microorganisms very relevant in the food industry and are relatively safe (GRAS) (generally regarded as safe). (José Miguel Miquelão Santos, 2021) . For that reason, they represent a valuable tool in biotechnology, for example in the pharmaceutical industry, in the plasmid DNA (pDNA) production, in DNA vaccines or in the expression of recombinant proteins for mucosal vaccination. *L. lactis* has also been shown to not colonise the intestine, so it is considered a viable method for drug delivery. (Alexander Bolotin, 2001)

However, the yields of recombinant proteins produced by *L. lactis* are disappointing and undermine their applicability (José Miguel Miquelão Santos, 2021). The *nth* gene, an homologue to the Endonuclease lll (Endo lll) of E. coli, has been identified in *L. lactis* in 2001 to be a potential target for deletion for increasing the yield of recombinant proteins. (Alexander Bolotin, 2001) . Endo lll is a DNA glycosylase, a repair enzyme with redundant activity and its deletion could decrease the non-specific digestion of DNA. (José Miguel Miquelão Santos, 2021)

Classic phylogenetic methodologies can be combined with Gene Neighbourhood Analysis to uncover new insights on the evolution of genes and of gene families. Gene neighbourhood analysis is a subfield of Comparative Genomics, consisting in the analysis of synteny shared by genome regions where the genes under analysis reside (Cristina G. Ghiurcuta, 2014). Regarding our gene of interest, nth1, the analysis of synteny will allow to identify rearrangements and duplications responsible for the dispersion of the members of the Nth gene family throughout the genome of the Lactobacillales species under analysis. (Cristina G. Ghiurcuta, 2014).

Although the identification of the orthologue status is important to unravel the evolutionary history of a gene family, there are other events driving the evolution of genes and genomes, for example gene duplications, horizontal gene transfer or gene loss and gain. (Paulo Jorge Dias, 2017)

An horizontal gene transfer (HGT) stands for the acquisition of DNA horizontally. This phenomenon is prevalent in prokaryotes, where it is one of the main mechanisms contributing to genetic variation and genome evolution. HGT can occur by the cellular uptake of exogenous DNA (transformation), the movement of chromosomal DNA via viruses (transduction) and DNA transfer via cell-cell contact (conjugation), and others (Rebecca J. Hall, 2020). Gene duplication is one of the most important forces driving the evolution of genetic functional innovation. After gene duplication, many scenarios can take course depending on their function, mode of duplication, expression rate and the organism taxonomic lineage. Some os these scenarios are the inactivation of one of the copies (pseudogenization), the maintenance of the two copies (dosage effect), the adoption of part of the function or of the expression pattern of their parental gene (subfunctionalization), or the acquisition of a related or new function (neofunctionalization). (Cláudia P. Godinho, 2018)

The purpose of this project is to reconstruct the evolution of the *nth* gene and potentially identify related endonucleases in an attempt to understand how Nth affects the pDNA production, improving the corresponding yields.

Methods and Results

Identification of the Nth homologues encoded in the genomes of the Lactobacillales species

This project is based on two databases developed in-house by the BSRG group, the Genome DB and the Blast DB. The Genome DB compiles the genomic and biological information on 557 Lactobacillales strains (corresponding to 256 different species), comprising a total of 1,225 million genes. The source of this information was the reference genome database and in the case of the Lactococcus lactis strain LMG 19460, whose genome was recently sequenced using an ONT- based long read sequencing approach, it was necessary to perform the corresponding genome annotation using Prokka software suite, since this genome has yet to be deposited in GenBank (this work was not done in this project).

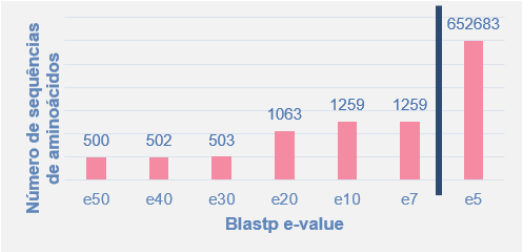


Figure 1- Number of sequences retrieved after constraining and traversing the pairwise similarity network at different e-values using the Nth sequence from *L. lactis* IL1403 as 'starting node'. QL queries to the Genome and Blast DB allow to traverse a network representing amino acid sequence similarity at different e-value thresholds, ranging from E-50 to E-5, using the starting node the L. lactis IL1403 protein. These queries were done over the interface provided by the SQLite Browser software suite. This analysis identified three successive plateaus, depending on the e-value threshold applied. An e-value of E-10 was chosen to constrain the pairwise similarity network and 1259 amino acid sequences of Nth homologues were retained for further evolutionary analyses.

Multiple Alignment of the amino acid sequences of the Nth homologs, Tree construction and Phylogenetic analysis

After the retrieval of the Nth homolog, the 1259 amino acid sequences were aligned using Muscle software suite (Figure 2).

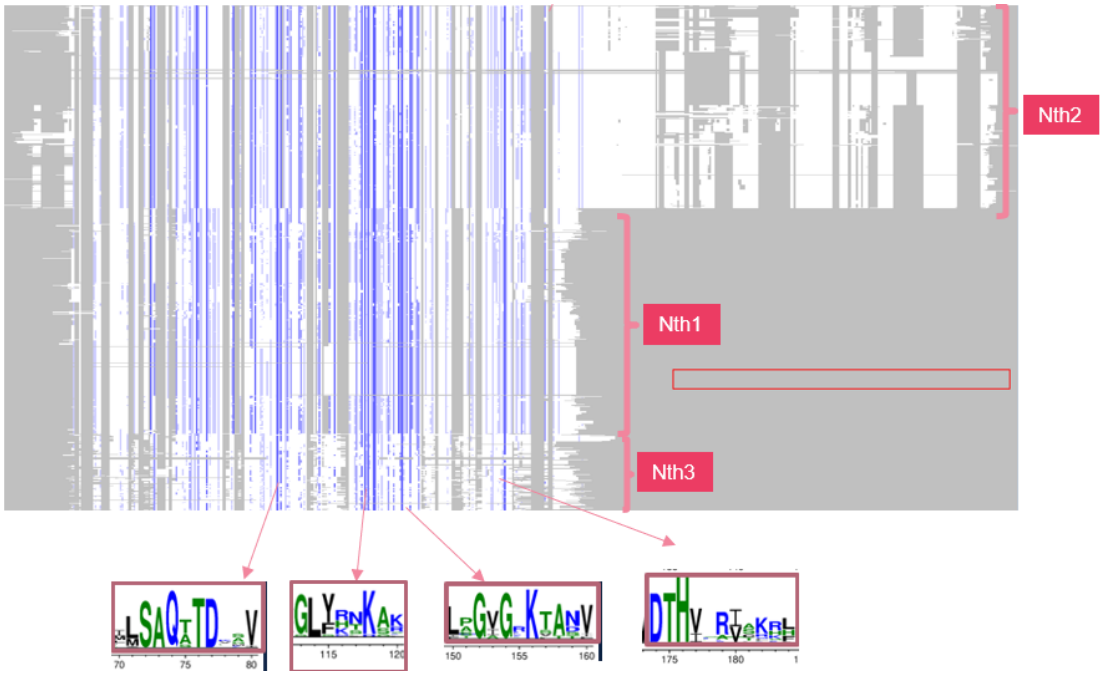


Figure 2- Overview of the multiple alignment, generated by Muscle, of the three endonucleases identified in this project. The intensity of the colour blue represents the conservation of the aligned amino acid sequences. The size of the amino acid in the sequence logos obtained using the LogOddsLogo software suite represent the corresponding frequency in each column of the multiple alignment. The conserved amino acid regions are augmented with the respective sequence Logos to increase the comprehension of the gathered results.

The analysis of this multiple alignment allowed the identification of three distinct regions. One region the aligned amino acid sequences of a subset of the Nth homologs, contains the aligned amino acid sequences of a subset of the Nth homologs, selected to be the main focus of this study, henceforth referred to as Nth1. The other two regions contained the aligned sequences of two additional endonucleases henceforth referred to as Nth2 and Nth3. Subsequently, the Protdist and Neighbour algorithms made available by the PHYLIP software suite were used to construct a global phylogenetic tree representing the diversity of the members of the Nth gene family (Figure 3). After analysing the phylogenetic tree represented in the Figure 2, the identification of the members in each Nth revealed that; the taxonomic family Leuconostocaceae doesn't have the *nth1* gene, although is in the presence of both other Nth's identified in this research, *nth2* and *nth3* and *Lactococcus lactis* lacks the *nth3*, but is in the

presence of the *nth1* and *nth2* genes*,* where the *nth2* gene represents the *mutY* present in E. coli.

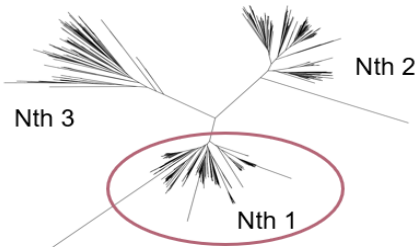


Figure 3- Global phylogenetic tree representing the diversity of the members of the Nth gene family.

Phylogenetic Analysis of the Nth1 gene subfamily

Correction of the DNA and Amino acid sequences of the Nth1 homologs and construction of the corresponding phylogenetic tree

This project was focused on the *L. lactis nth* gene. For that reason, the initial gene and protein dataset comprising 1259 members of the the Nth gene family was reduced to 568 genes, corresponding to the Nth1 homologs. The corresponding amino acid sequences were used ot construct a preliminary tree (Figure 4). Five Nth1 homologs were identified to be problematic because of the discrepancy in distance within the corresponding phylogenetic branches. The DNA and protein sequences of four of these Nth1 homologs were corrected since they presented errors (most likely, resulting from poor-quality sequencing and/or assembly of specific genome regions in the corresponding genome projects). This correction was done by comparing close genes defined by the phylogenetic analysis. One of these sequences was eliminated, presumed to be a false positive Nth1 homolog. This sequence consisted in a protein fragment, showing a poor alignment with the Nth1 homolog selected as reference (this alignment is represented in Annex 3).

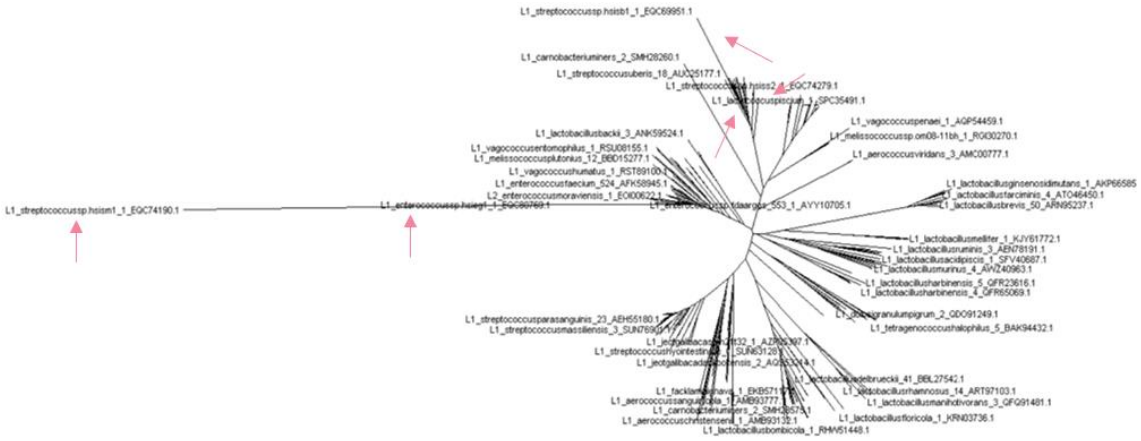


Figure 4-Preliminary phylogenetic tree obtained with MUSCLE Software and PHYLIP Software, based on ProTdist and Neighbour algorithm. Marked with the pink arrows are the sequenced identified has problematic and the target of this sequence correction, which correspond to L1_streptococcussp. hsism1_1_EQC74190.1, L1_enterococcussp. hsieg1_1_EQC80769.1, L1_streptococcussp. hsisb1_1_EQC69952.1, L1_streptococcusparauberis_7_AEF25426.1 and L1_streptococcussp. hsisb1_1_EQC69951.1.

After the sequence correction of this handful of cases a more refined phylogenetic tree was constructed using the bayesian algorithm made available by MrBayes software suite (Figure 5). The End lll/Nth1 phylogenetic tree is divided into 17 clusters. These clusters were proposed based on 1) the distances and topology of the phylogenetic tree, 2) the posterior probability of each ramification and 3) the taxonomic origin of the Nth1 homologs comprised in each cluster. Hence, each cluster was labelled based on the associated Lactobacillales taxonomic families.

In the upper part of the tree, represented in detail in Figure 6, is divided into 9 clusters. These clusters comprised genes encoded in species classified in the following the taxonomic families Carnobactereaceae, Lactobacillaceae, Enterococcaceae, Aerococacceae and Streptococcaceae, these clusters are labelled L1, L2, S1, EA1, C3, A1, L3, C1 and A2. The ORF *Lactobacillus ginsenosidimutans* EMML 3041, was chosen as outgroup due to the strong dissimilarity of its amino acid sequence when compared with those of the remaining members of the *nth1* gene subfamily. This ORF is localized in cluster L3. Close to the root of the tree are the clusters, L1 and L2. The cluster EA1 comprises genes encoded in species classified in two different taxonomic families, three strains of the same species *Tetragenococcus* and two *Aerococccus* strains. The cluster S1, contains only part of the *Streptococcus* species and strains analysed in this work.

The cluster A1 comprises genes encoded only in *Aerococcus* species. It is also the only cluster comprising more than one strain of the same species residing in the same cluster (this observation is verified in Annex 6).

The middle part of this tree comprises genes encoded in species classified in the Carnobactereaceae and Enterococcaceae taxonomic families. These clusters are labelled C4, C5, C6, E2, E3 and E4, and are represented in detail in Annex 4.

In the lower part of the tree are residues all the Nth1 homologs encoded in the *Lactococcus* strains, in cluster S2. This cluster also comprises the genes encoded in the remaining *Streptococcus* species and strains (those not residing in cluster S1). The cluster S2 is shown in Figure 7, where is possible to verify that the Nth1 homologs encoded in the different *Lactococcus* species present a greater similarity to each other, the same occurring between the Streptococcus species (i.e. the two genera group into distinct subclusters). Also represented at the lower part of the tree is the CA1 cluster, comprising genes encoded in three Aerococacceae species and six Carnobactereaceae species.

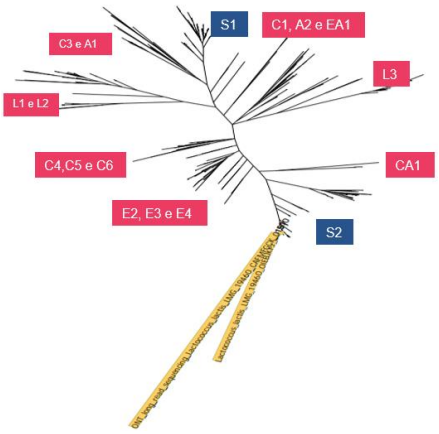


Figure 5- Phylogenetic Tree representing the diversity of the protein encoded in the *nth1* subfamily of genes of L. lactis. The tree was constructed using Muscle Software and MrBayes method.

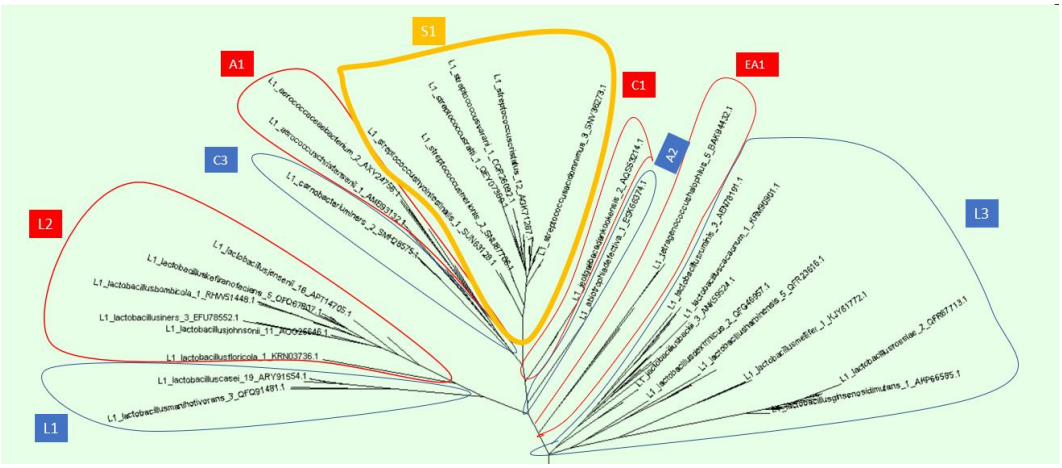


Figure 6- Upper part of the phylogenetic tree from Figure 5, showing the detailed composition of the clusters L1, L2, C1, C3, A1, A2, EA1 and S1 represented.

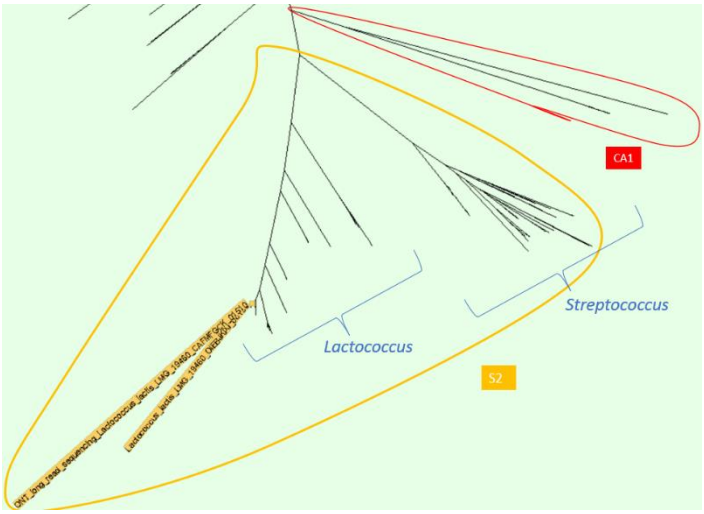


Figure 7- Lower part of the phylogenetic tree from Figure 5, showing in detail the composition of the cluster CA1 and S2 represented. In blue is represented the taxonomic families of the cluster S2. In the yellow boxes are represented the Nth1 homolog encoded in the LMG 19460 strain that in the main goal of this research project, referred to an ONT_long_read_sequencing_Lactococcus_lactis_LMG_CAFMFGCK_01510, sequenced by Nanopore e Lactococcus_lactis_LMG_19460_OIEBJKPJ_00150, from GenBank sequenced by *Illumina.*

Identification of conserved regions

By the analysis of Figure 2 is possible to identify four conserved regions, common to Nth1, Nth2 and Nth 3 and one region common to the Nth1 and Nth2.

The structural analysis of the Endo lll in E. coli sustains that this protein is a two-domain-helical protein, with one domain organized into a six-helix barrel and another domain containing an iron-sulfur $[4Fe\text{-}4S]^{2+}$ cluster (FeS). In E. coli four domains were found to be responsible for the DNA distortion of the Endo lll in E. coli when compared to the wild type. (Structural basis for recognition and repair of the endogenous mutagen 8-oxoguanine in DNA, 2000)

When compared to the Nth1, the team identified the same four motifs responsible for the DNA distortion although the Nth1 lacks the domain containing the iron-sulfur $[4Fe\text{-}4S]^{2+}$ and cluster (FeS), this can be verified by the lack of four consecutive cystines in the multiple alignment, shown in Annex 4. All four motifs were found to be further in the Nth1 sequence, when compared to the End lll in E. coli. (Structural basis for recognition and repair of the endogenous mutagen 8-oxoguanine in DNA, 2000)

The first motif one being from Ser^{70}-Asn^{81}. In E. coli this conserved region is responsible for the connection of the alpha-B and alpha-C domains, which will affect the secondary structure on Nth1. The domain from Ser^{70}-Asp^{77} is responsible for the DNA distortion, it is possible to observe that within the motif identified from Ser^{70}-Asn^{81}, the domain responsible for the DNA distortion. This domain is inserted into the minor groove of the DNA double helix to push the damaged nucleotide out. The void left after the eversion is filled by Gln^{74}, and the adjacent bases are stacked against its side chain amide group (the 3-base) and the planar peptide bond Gln^{74}-Cys^{75}. (Nikita A. Kuznetsov, 2015)

The second motif, from Gly^{114}-Lys^{120}, appears to be a highly conserved region in all *nth* genes and stands for the αF-helical- domain and is also responsible for the DNA distortion. The helices αF and the beginning of αG buttress the orphaned nucleotide opposite the lesion and two flanking nucleotides and introduce a kink into this strand. The third motif represented in Figure 2, Gly^{153}-Val^{160}, is found to be responsible for the αL-helical-domain, this helix is the only one out of the α-helices that actually contacts the DNA backbone, explaining why this is a highly conserved region of the HhH-helix-hairpin. The fourth motif, Asp^{173}-Lys^{190}, stands for in the literature for the α-M helix. The amino acid Arg^{179} is found to be responsible the DNA distortion. (Nikita A. Kuznetsov, 2015)

Comparative Genomics

Synteny was evaluated in this research with the analysis of the neighbourhood, up to 15 genes to the left, and 15 genes to the right of the *nth1* homologs in our Genome DB with the Software Cytoscape.

By the comparative genomics analysis, we have identified no poor synteny between each cluster all around the tree. We focused on this analysis in the S1 and S2 clusters, that unexpectedly were in opposite sides of the phylogenetic tree, after the comparative genomics analysis we concluded that the synteny within the two clusters was strong, they were in a presence of ortholog genes, but between S1 and S2 there wasn't a strong synteny, all thought there we some connections between the neighbourhood.

Discussion

After analysing the phylogenetic tree represented in the Figure 2, the identification of the members in each Nth revealed that the taxonomic family Leuconostocaceae doesn't have the *nth1* gene, but it is in the presence of both other Nth's identified in this research, *nth2* and *nth3*. This Family is the oldest in the Lactobacilalles order, which can lead to the conclusion that the End lll has surged later in the evolution. Because the taxonomic classification of the Lactobacillales order was reviewed in 1995, this hypothesis might not be accurate. Another possible explanation is the deletion of the nth1 gene from the root of the Leuconostocaceae

branch, this might indicate that the nth1 is not essential to the organism, which has been proven by the deletion of the *nth1* in *L. lactis*.

The domains responsible for the DNA distortion being highly conserved leads to the conclusion that there are essential to the Nth catalysis and DNA binding. The FeS domain, in E. coli responsible for the catalysis of the DNA binding reaction, being eliminated from Nth1 reveals that this domain is not essential to the Nth function.

From the analysis of Figure 5, two clusters were identified to have more than one taxonomic family, CA1 and EA1. One possible justification for these clusters is horizontal transfer of genes, which can be justified by the comparative genomics analysis.

The cluster EA1 composed by Enterococcaceae and Aerococacceae, two families which are not close related, from the comparative genomic analysis, the team observed that there was a strong synteny within this cluster. A possible explanation for this proximity in *nth1* evolution, is an horizontal transfer of gene, it is not possible to determine the receptor and donor families in this cluster. In the CA1 cluster, another example of horizontal transfer of genes, is possible to consider Carnobactereacea family has the donor, for being the most frequent family in this cluster.

In general, there is only one *nth1* present in a species, except in species of the taxonomic families Aerococacceae and Carnobactereacea, these are also families that have shown to have examples of horizontal transfer of genes. The hypothesis is that in case of a horizontal transfer of genes or local duplication of the gene *nth1* the species was the presence of two *nth1* genes, but during the evolution these have been lost, showing that the possible gene duplication in cluster A1 must have happened recently because most of the strains have two *nth1* genes present. This hypothesis sustains the idea that there is only necessary one nth1 for the function to be ideal, although the presence of two *nth1* genes does not compromise cell survival.

In Figures 3 we see that the taxonomic Family Aerococacceae appears represented across the tree, what was not expected, the taxonomic families should be close by, seeing they are more related within the family, in the *nth1* evolution, that does not happen. Which can be explained by an horizontal transfer of genes in the root of the tree, which cannot be traced anymore.

By the comparative genomics analysis, we have concluded that there was no strong synteny between each cluster all around the tree. We focused on this analysis in the S1 and S2 clusters, that unexpectedly were in opposite sides of the phylogenetic tree, after the comparative genomics analysis we concluded that the synteny within the two clusters was strong, they were in a presence of ortholog genes, but between S1 and S2 there wasn't a strong synteny, which was expected by the phylogenetic analysis, seeing that S2 and S1 are in opposite points of the phylogenetic tree. This analysis sustains two possible theories. First one, the S1 and S2 common ancestors had the same nth1 gene that duplicated, and a part of the *Streptococcus* retained one of this nth1 gene and the *Lactococcus* and a part of the *Streptococcus* retained another *nth1* gene. This theory would have been proven if in the comparative genomics the synteny between the S1 and S2 was stronger than the synteny between S1 and the other clusters and the synteny of S2 between the other clusters. The second theory sustains that S1

and S2, don't share the same common ancestor, but share an older ancestor, now untraceable, the S1 and S2 genes were shared to other families and diverged throughout the tree. S1 shared it's nth1 gene with the clusters at the top of the tree, S2 with the clusters at the bottom of the tree and either S1 or S2 has shared the nth1 with the clusters at the middle of the tree. This theory could have been proven if in the comparative genomics the synteny of the S2 and S1 clusters was stronger with the other clusters nearby then the synteny between them.

In conclusion we can that the comparative genomics and phylogenetic analysis are in agreement. With this project we have identified one potential target to eliminate in *L. lactis,* the *nth2* gene. Other approach could be the use of a Leuconostocaceae species, seeing they already don't have the nth1 gene. Instead of eliminating the nth1 gene, which has been proven to be a difficult and laborious process, the compromise of one of the conserved and apparently essential regions, could be a viable approach.

References

Alexander Bolotin, P. W. (2001, February 5). The Complete Genome Sequence of the Lactic Acid Bacterium Lactococcus lactis ssp. lactis IL1403.

Cláudia P. Godinho, P. J.-C. (2018, October 15). The Paralogous Genes PDR18 and SNQ2, Encoding Multidrug Resistance ABC Transporters, Derive From a Recent Duplication Event, PDR18 Being Specific to the Saccharomyces Genus.

Cristina G. Ghiurcuta, B. M. (2014, Bioinformatics, Volume 30, Issue 12, 15 June 2014, Pages i9–i18,). *Evaluating synteny for improved comparative studies* (Vol. 30).

José Miguel Miquelão Santos, G. A. (2021, March 16th). One-plasmid based CRISPR-Cas9 editing for gene deletion in Lactococcus latis .

Nikita A. Kuznetsov, O. A. (2015, October 22). Conformational Dynamics of DNA Repair by Escherichia coli.

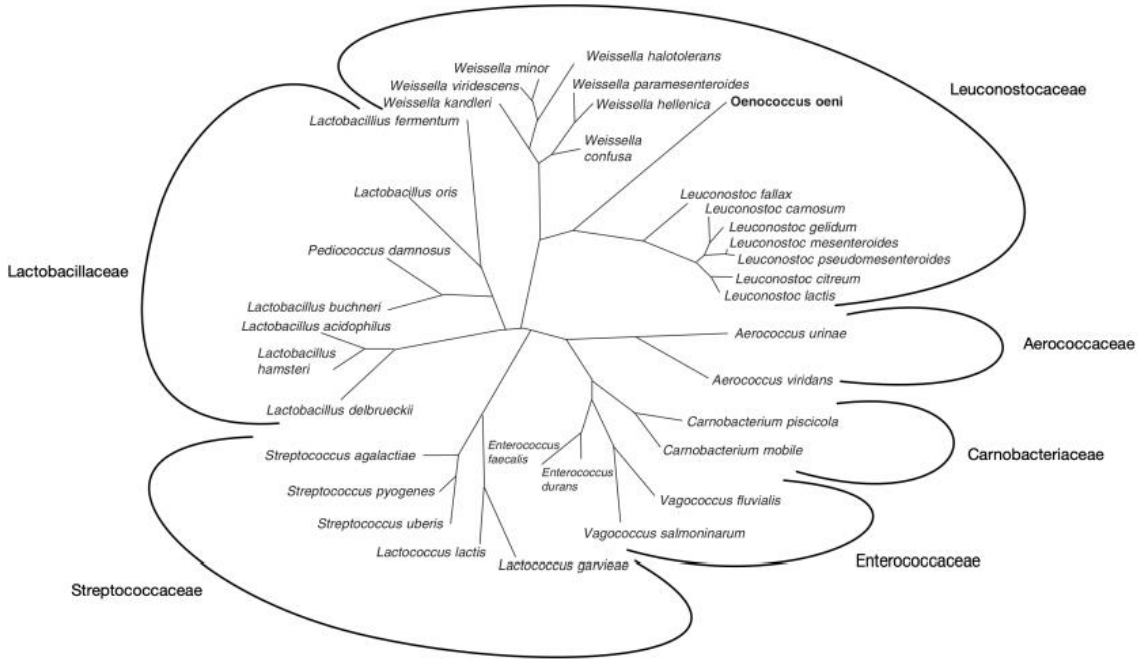
Paulo Jorge Dias, M. P. (2017). The Zygosaccharomyces bailii transcription factor Haa1 is required for acetic acid and copper stress responses suggesting subfunctionalization of the ancestral bifunctional protein Haa1/Cup2.

Rebecca J. Hall, F. J.-S. (2020, July 17). Horizontal Gene Transfer as a Source of Conflict and Cooperation in Prokaryotes.

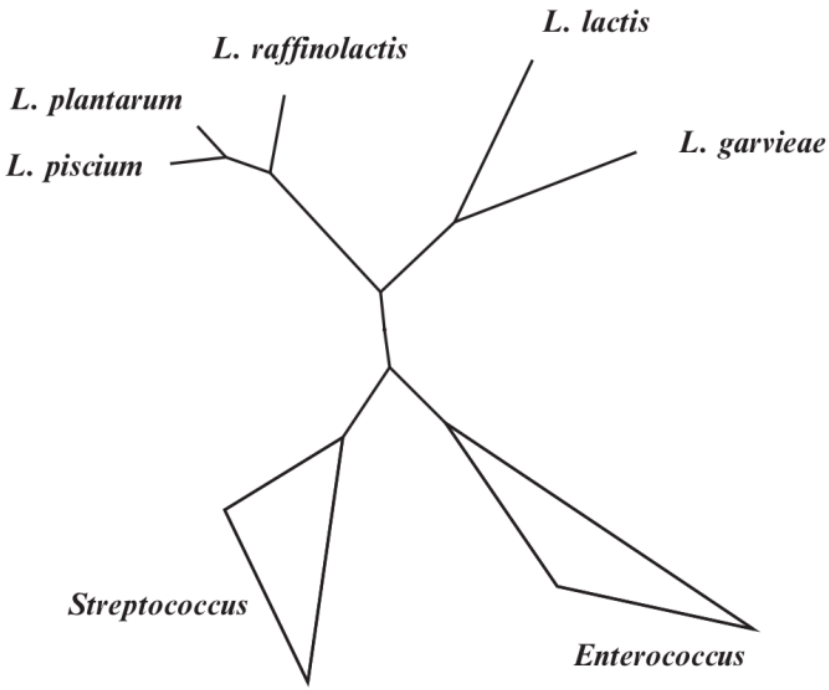
Structural basis for recognition and repair of the endogenous mutagen 8-oxoguanine in DNA. (2000, February 24).

Y Saito, F. U. (1997, Jun). Characterization of endonuclease III (nth) and endonuclease VIII (nei) mutants of Escherichia coli K-12. doi:10.1128/jb.179.11.3783-3785.1997

Annexes



Annex 1- Unrooted phylogenetic tree showing the relationship of Oenococcus oeni and lactic acid bacteria based on 16S rRNA sequences. (Modified and reprinted from Dicks et al., 1995. International Journal of Systematic Bacteriology 45: 395–397).



Annex 2-Phylogenetic tree of the species of the genus Lactococcus. The tree was constructed by a maximum-likelihood analysis of more than 50,000 full 16S rRNA sequences. The bar indicates 5% estimated sequence differences (courtesy of Wolfgang Ludwig, TU Munich, Germany).

>L1_enterococcussp.hsieg1_1_EQC80769.1_upstream_sequence
AGTCTTTTCAAGATCACGCCAACGTTTTAAGAAGTACAAGCTTCACTTCAGGTAAAGCACATTATAGTGGGGTCTGGAAAAGTCAAACGACTTCTTTTTACGATGCGACCAGCTATCTGACAGGACGC
TATGCCACTGATCCAGGTTACGCACAAAAATTAAACTGGCTGATTGAAACGTATCAACTAACTCGTTTTGATACACCAAGTACTGGTGCCACAACGACTTCTGTAACGACAACGACAACGCAATCCAC
ATCGACGACAACAACGACACAAAGTACGGGACAAAGTTATGTCGTAGCATCTGGCGACAGTCTTTGGGATATTGCTTCACGTTTTGGAACAACGGTGGATCAAATCATGGCTGCCAATGGATTAAGCA
GTGACTTGATCGTCGTAGGACAGGAATTGATCGTTTGATTTTCGCTATTTATTAGCTAAACGCCAAATGAAATTTCACAAAAACTTTTGTTTTTGTTCTCTTTTAGTAACTAAAAGAAGCAAAGCAAA
AGTTTTTTTCTTTGTGAAAAAAACATAAATATTTGGAAAACGCTAACTTTTTGTTCGCAAGTCTGCCGTTTTTTGGGTAGAATAGGAATATGAACTGGAGGGATTGCAAATGACAAAGATTTATCAAT
CTGTTACCGAGCTAGTTGGGCATACGCCCATCGTCAAATTAAACAAGGTGGTTGGAGAGGATTCCGCTGATGTCTACGTTAAACTTGAGTTTTATAATCCAGGGAGCAGCGTGAAGGACCGGATCGCT
TTGGCTATGATCGAAAAAGCCGAAAAAGACGGCTCTTTGAAAGCAGGCAATACGATCGTGGAACCAACTTCCGGTAATACTGGGATCGGCTTGGCGATGATCGGTGCAGCGAAAGGGTACCGAGTGGT
GATCGTGATGCCTGAGACCATGAGTATCGAACGGCGCAAGCTGATGCAAGCCTATGGTGCTGAATTGATTTTGACACCAGGCTCAGAAGGAATCACAGGTTCAATCAATAAAGCCAAGGAATTGGCTG
AACAAGAGGGCTATTTTATGCCGCTGCAATTTGAAAATCCGGCAAATCCCGAGATTCACGAGAAAACAACTGGAAAAGAAATAGAGAATGCCTTTGAATCAAAAGGATTAGATGCCTTTGTCGCTGGT
ATCGGTACAGGCGGCACGATCACTGGTGCGGGTCGAGAATTAAAACGAGTGTATCCAGACTTAAAAATCTATGGCGTTGAACCGGCAGAATCGGCTATCTTAGAAGGCGGACAACCTGGTCCTCACAA
AATCCAAGGAATTGGAACCGGCTTTGTACCTAAAGTTCTTGATACAGAGATCTTTGATCAAGCAATTGCCGTATCAAGTGAAGATGCGATGGCAACAGCGAGAGAAGTCGGCCGTAAAGAGGGCTTGC
TCGTAGGTATTTCTTCTGGAGCAGCGATCCATGCAGCACTCAAAGTTGCCAAAGAACTGGGAAAAGGCAAAAAAGTCTTAGCGATCGTCCCGGACAATGGAGAACGCTATCTATCGACTGCTTTATAT
CAAGTGGAAGAATAACAAGGAAAGCGGGCATTGGCAGAATGCCTGCTTTTTCTCCCCTTTTCTTAAGGACTGTGGAATTTTTCGTGTTTGTTTTGTGGTTGATTAAGAATTTTTAGAGAAAGCTGTTT
GAATTTTTACCATCTACATAGTATGATATTACAGATAAGAGAATTGGAGGAACTCCTGTGAACACGATATATGATGAGGCAATCAAACAATTAAAACAAGAAAATATCCGTATCACTCCACAAAGGGT
TGCGATACTTGAGTTTTTGGCTAGCCACGCATCACATCCTACCGCAGAAGAAATTTACCGAGCGATCGAAGTCCACTTTCCCGGTATTAGTGTAGCGACAGTCTACAATAATCTGCGACTGTTTACGG
AGATCGGTTTTGTGAAAGAAATGAACTACGGCGATGCATCCAGTCGCTTTGACTTTAAGACAGAGCCGCATTATCATGCAATTTGTCAACGTTGCGGCAAGATTTCTGATTTTTACTATCCTGGATTG
GAAGATGTCGAAATGGCCGCTTCCCAACTGACCGGATATAAAATCCAAAGCCATCGCTTAGAAGTATATGGATTGTGTCCAGAGTGTAAAGGAGAAGAGGAATGAAAATTATTTCAACGGATAAAGCA
CCGCAAGCAATTGGTCCTTATGTACAAGCACGGATTGCTAATGGTTTCTTGTTTGCTTCCGGACAAGTGCCGTTAGATCCTGTTACAGGAGAAGTTGTGGGAGCAACTATCGAAGTTCAGACACAGCA
AGTTCTCAAAAATATCGAGGCAATTTTAACTGCCGCCGGTATTACACCAGATGAAATCGTCAAAACAACATGTTTCTTAAAAAACATGGATGATTTTGTTCCATTCAATAATACATACAGCACACTTT
TTCATGAACATTTGCCTGCACGTTCAGCAGTTGAAGTGGCCCGTTTACCAAAGGATGTATTGGTGGAAGTAGAGATCATTGCTGTGGTGGCACAATAAAGGAGTCAATGGGACGGTCGTATATATGCG
CAGTCCCATTTTTGTTGAGAATGGATTTTATTCTTGATGATGCCAAAATAAAATTGAAAATGGTTTTTATGTAACGATGGGAAAAGAGGACAATAAGATGATTTCTTTACAAGAATATTTAGCCCAGG
GGCAAACAACTGTCTCCAATCTGCTTTTGGAAAACTATCATAGCCTTGGTTTAACCAATGATGAATTTGTATTATGGCTTCAGTTGTATCGTTACCACGAACAAGGCAATGATTTTCCAGATTTAGCG
ATGATTGCTAAAGGGATGGGAAGTGAGCAAAAAGAAATCTATAAATGGTTGAATGAGCTATTGAATAAACAGATCATCGCCATCGAATCCAAGAAAGATTCAGCAGGGAAGATGGTGGATTACTATGA
GTTTTCTGGAATCTATGAACGACTGGAAGTTTGTTTGCGGCAAAAAAATCAGCAGGAGGAAAAGCAGAGCTATGATTCAAAAGTACGTTCGCTTTATCAAATGTTTGAGTCCGAATTTGGCCGTCCGC
TTTCAGCAATCGAATACCAGCGAATATCACAATGGCTGGAAGAAGATAAATACCAGCCGGATTTGATTCAATTTGCGTTACGTGAGGCGGTTTTGAATCAAGCCTACAGTTTGAATTATATTGATCGC
ATTTTACTTTCTTGGGAACGGAAAAATATTACCTCGAAAGAACAAGTAGAAGAAGAACAAAAACGCCGTAAAAAACAGTTTATGCAGAAAGAAAATTTGGAACAAACGAATGGACCGGTTCCTAAGAT
TCCTATGCATAATTGGTTAGAAGGGGAATAAAGGAGAGTTGCCATGATCAATAAAGCAAAAACTATGATCGCATTGGAACAAATGTACCAAATGTTCCCCGACGCTCACGGAGAACTGATTTCCAAAA
ATCCGTTTGAACTGTTGATTGCTGTGATTTTAAGTGCTCAGGCTACGGATGTGTCGGTCAATAAAGTGACCCCCACACTATTTGCTGCCTATCCGACACCTGAGGCCTTGGCAGCAGCTCCTGTTGAG
GAAATCATTGAGAAAATACGGACGATCGGTCTTTATCGCAACAAAGCAAAAAATATCAAGGCTTGTGCGTCCCAGCTTATCGAACGATTCAATGGCCAAGTACCTAGAACCCGCGAGGAATTAGTGTC
ATTACCGGGGGTCGGAAGAAAAACAGCCAACGTTGTTTTGGGCGATGCGTTTGGTATTCCGGCAATCGCTGTGGATACCCACGTCGAACGAGTAACCAAGCGCTTACGGATTTGCCGATTAGATGCAA
ATGTCTTAGAAGTAGAACAAACATTGATGAAAAAAGTGCCAGAAGATCTTTGGGTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNATTG

>L1_enterococcussp.hsieg1_1_EQC80769.1_DNA_coding_sequence
ATGAAAAAAGTGCCAGAAGATCTTTGGGTCAAAACACATCACACACTGATTTTCTTTGGACGATACCATTGCACAGCTAGAGCGCCAAAATGTGAAGTTTGCCCATTATTGACAATGTGCCAAGAAGG
ACAACTGCGGATGCGCGCAAAAAAAGTTCAGCCACCTCAATGA

>L1_enterococcussp.hsieg1_1_EQC80769.1_upstream_and_DNA_coding_sequence
AGTCTTTTCAAGATCACGCCAACGTTTTAAGAAGTACAAGCTTCACTTCAGGTAAAGCACATTATAGTGGGGTCTGGAAAAGTCAAACGACTTCTTTTTACGATGCGACCAGCTATCTGACAGGACGC
TATGCCACTGATCCAGGTTACGCACAAAAATTAAACTGGCTGATTGAAACGTATCAACTAACTCGTTTTGATACACCAAGTACTGGTGCCACAACGACTTCTGTAACGACAACGACAACGCAATCCAC
ATCGACGACAACAACGACACAAAGTACGGGACAAAGTTATGTCGTAGCATCTGGCGACAGTCTTTGGGATATTGCTTCACGTTTTGGAACAACGGTGGATCAAATCATGGCTGCCAATGGATTAAGCA
GTGACTTGATCGTCGTAGGACAGGAATTGATCGTTTGATTTTCGCTATTTATTAGCTAAACGCCAAATGAAATTTCACAAAAACTTTTGTTTTTGTTCTCTTTTAGTAACTAAAAGAAGCAAAGCAAA
AGTTTTTTTCTTTGTGAAAAAAACATAAATATTTGGAAAACGCTAACTTTTTGTTCGCAAGTCTGCCGTTTTTTGGGTAGAATAGGAATATGAACTGGAGGGATTGCAAATGACAAAGATTTATCAAT
CTGTTACCGAGCTAGTTGGGCATACGCCCATCGTCAAATTAAACAAGGTGGTTGGAGAGGATTCCGCTGATGTCTACGTTAAACTTGAGTTTTATAATCCAGGGAGCAGCGTGAAGGACCGGATCGCT
TTGGCTATGATCGAAAAAGCCGAAAAAGACGGCTCTTTGAAAGCAGGCAATACGATCGTGGAACCAACTTCCGGTAATACTGGGATCGGCTTGGCGATGATCGGTGCAGCGAAAGGGTACCGAGTGGT
GATCGTGATGCCTGAGACCATGAGTATCGAACGGCGCAAGCTGATGCAAGCCTATGGTGCTGAATTGATTTTGACACCAGGCTCAGAAGGAATCACAGGTTCAATCAATAAAGCCAAGGAATTGGCTG
AACAAGAGGGCTATTTTATGCCGCTGCAATTTGAAAATCCGGCAAATCCCGAGATTCACGAGAAAACAACTGGAAAAGAAATAGAGAATGCCTTTGAATCAAAAGGATTAGATGCCTTTGTCGCTGGT
ATCGGTACAGGCGGCACGATCACTGGTGCGGGTCGAGAATTAAAACGAGTGTATCCAGACTTAAAAATCTATGGCGTTGAACCGGCAGAATCGGCTATCTTAGAAGGCGGACAACCTGGTCCTCACAA
AATCCAAGGAATTGGAACCGGCTTTGTACCTAAAGTTCTTGATACAGAGATCTTTGATCAAGCAATTGCCGTATCAAGTGAAGATGCGATGGCAACAGCGAGAGAAGTCGGCCGTAAAGAGGGCTTGC
TCGTAGGTATTTCTTCTGGAGCAGCGATCCATGCAGCACTCAAAGTTGCCAAAGAACTGGGAAAAGGCAAAAAAGTCTTAGCGATCGTCCCGGACAATGGAGAACGCTATCTATCGACTGCTTTATAT
CAAGTGGAAGAATAACAAGGAAAGCGGGCATTGGCAGAATGCCTGCTTTTTCTCCCCTTTTCTTAAGGACTGTGGAATTTTTCGTGTTTGTTTTGTGGTTGATTAAGAATTTTTAGAGAAAGCTGTTT
GAATTTTTACCATCTACATAGTATGATATTACAGATAAGAGAATTGGAGGAACTCCTGTGAACACGATATATGATGAGGCAATCAAACAATTAAAACAAGAAAATATCCGTATCACTCCACAAAGGGT
TGCGATACTTGAGTTTTTGGCTAGCCACGCATCACATCCTACCGCAGAAGAAATTTACCGAGCGATCGAAGTCCACTTTCCCGGTATTAGTGTAGCGACAGTCTACAATAATCTGCGACTGTTTACGG
AGATCGGTTTTGTGAAAGAAATGAACTACGGCGATGCATCCAGTCGCTTTGACTTTAAGACAGAGCCGCATTATCATGCAATTTGTCAACGTTGCGGCAAGATTTCTGATTTTTACTATCCTGGATTG
GAAGATGTCGAAATGGCCGCTTCCCAACTGACCGGATATAAAATCCAAAGCCATCGCTTAGAAGTATATGGATTGTGTCCAGAGTGTAAAGGAGAAGAGGAATGAAAATTATTTCAACGGATAAAGCA
CCGCAAGCAATTGGTCCTTATGTACAAGCACGGATTGCTAATGGTTTCTTGTTTGCTTCCGGACAAGTGCCGTTAGATCCTGTTACAGGAGAAGTTGTGGGAGCAACTATCGAAGTTCAGACACAGCA
AGTTCTCAAAAATATCGAGGCAATTTTAACTGCCGCCGGTATTACACCAGATGAAATCGTCAAAACAACATGTTTCTTAAAAAACATGGATGATTTTGTTCCATTCAATAATACATACAGCACACTTT
TTCATGAACATTTGCCTGCACGTTCAGCAGTTGAAGTGGCCCGTTTACCAAAGGATGTATTGGTGGAAGTAGAGATCATTGCTGTGGTGGCACAATAAAGGAGTCAATGGGACGGTCGTATATATGCG
CAGTCCCATTTTTGTTGAGAATGGATTTTATTCTTGATGATGCCAAAATAAAATTGAAAATGGTTTTTATGTAACGATGGGAAAAGAGGACAATAAGATGATTTCTTTACAAGAATATTTAGCCCAGG
GGCAAACAACTGTCTCCAATCTGCTTTTGGAAAACTATCATAGCCTTGGTTTAACCAATGATGAATTTGTATTATGGCTTCAGTTGTATCGTTACCACGAACAAGGCAATGATTTTCCAGATTTAGCG
ATGATTGCTAAAGGGATGGGAAGTGAGCAAAAAGAAATCTATAAATGGTTGAATGAGCTATTGAATAAACAGATCATCGCCATCGAATCCAAGAAAGATTCAGCAGGGAAGATGGTGGATTACTATGA
GTTTTCTGGAATCTATGAACGACTGGAAGTTTGTTTGCGGCAAAAAAATCAGCAGGAGGAAAAGCAGAGCTATGATTCAAAAGTACGTTCGCTTTATCAAATGTTTGAGTCCGAATTTGGCCGTCCGC
TTTCAGCAATCGAATACCAGCGAATATCACAATGGCTGGAAGAAGATAAATACCAGCCGGATTTGATTCAATTTGCGTTACGTGAGGCGGTTTTGAATCAAGCCTACAGTTTGAATTATATTGATCGC
ATTTTACTTTCTTGGGAACGGAAAAATATTACCTCGAAAGAACAAGTAGAAGAAGAACAAAAACGCCGTAAAAAACAGTTTATGCAGAAAGAAAATTTGGAACAAACGAATGGACCGGTTCCTAAGAT
TCCTATGCATAATTGGTTAGAAGGGGAATAAAGGAGAGTTGCCATGATCAATAAAGCAAAAACTATGATCGCATTGGAACAAATGTACCAAATGTTCCCCGACGCTCACGGAGAACTGATTTCCAAAA
ATCCGTTTGAACTGTTGATTGCTGTGATTTTAAGTGCTCAGGCTACGGATGTGTCGGTCAATAAAGTGACCCCCACACTATTTGCTGCCTATCCGACACCTGAGGCCTTGGCAGCAGCTCCTGTTGAG
GAAATCATTGAGAAAATACGGACGATCGGTCTTTATCGCAACAAAGCAAAAAATATCAAGGCTTGTGCGTCCCAGCTTATCGAACGATTCAATGGCCAAGTACCTAGAACCCGCGAGGAATTAGTGTC
ATTACCGGGGGTCGGAAGAAAAACAGCCAACGTTGTTTTGGGCGATGCGTTTGGTATTCCGGCAATCGCTGTGGATACCCACGTCGAACGAGTAACCAAGCGCTTACGGATTTGCCGATTAGATGCAA
ATGTCTTAGAAGTAGAACAAACATTGATGAAAAAAGTGCCAGAAGATCTTTGGGTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNATTGATGAAAAAAGTGCCAGAAGATCTTTGGGTCAAAACACATCACACACTGATTTTCTTTGGACGATACCATTGCACAGCTAGAGCGCCAAAATGTGAAGTTTGCCCATTATTGACAATGTGC
CAAGAAGGACAACTGCGGATGCGCGCAAAAAAAGTTCAGCCACCTCAATGA

(First sequence correction- Removal of the sequencing mishaps)

>L1_enterococcussp.hsieg1_1_EQC80769.1_upstream_e_DNA_coding_sequence

AGTCTTTTCAAGATCACGCCAACGTTTTAAGAAGTACAAGCTTCACTTCAGGTAAAGCACATTATAGTGGGGTCTGGAAAAGTCAAACGACTTCTTTTTACGATGCGACCAGCTATCTGACAGGACGC
TATGCCACTGATCCAGGTTACGCACAAAAATTAAACTGGCTGATTGAAACGTATCAACTAACTCGTTTTGATACACCAAGTACTGGTGCCACAACGACTTCTGTAACGACAACGACAACGCAATCCAC
ATCGACGACAACAACGACACAAAGTACGGGACAAAGTTATGTCGTAGCATCTGGCGACAGTCTTTGGGATATTGCTTCACGTTTTGGAACAACGGTGGATCAAATCATGGCTGCCAATGGATTAAGCA
GTGACTTGATCGTCGTAGGACAGGAATTGATCGTTTGATTTTCGCTATTTATTAGCTAAACGCCAAATGAAATTTCACAAAAACTTTTGTTTTTGTTCTCTTTTAGTAACTAAAAGAAGCAAAGCAAA
AGTTTTTTTCTTTGTGAAAAAAACATAAATATTTGGAAAACGCTAACTTTTTGTTCGCAAGTCTGCCGTTTTTTGGGTAGAATAGGAATATGAACTGGAGGGATTGCAAATGACAAAGATTTATCAAT
CTGTTACCGAGCTAGTTGGGCATACGCCCATCGTCAAATTAAACAAGGTGGTTGGAGAGGATTCCGCTGATGTCTACGTTAAACTTGAGTTTTATAATCCAGGGAGCAGCGTGAAGGACCGGATCGCT
TTGGCTATGATCGAAAAAGCCGAAAAAGACGGCTCTTTGAAAGCAGGCAATACGATCGTGGAACCAACTTCCGGTAATACTGGGATCGGCTTGGCGATGATCGGTGCAGCGAAAGGGTACCGAGTGGT
GATCGTGATGCCTGAGACCATGAGTATCGAACGGCGCAAGCTGATGCAAGCCTATGGTGCTGAATTGATTTTGACACCAGGCTCAGAAGGAATCACAGGTTCAATCAATAAAGCCAAGGAATTGGCTG
AACAAGAGGGCTATTTTATGCCGCTGCAATTTGAAAATCCGGCAAATCCCGAGATTCACGAGAAAACAACTGGAAAAGAAATAGAGAATGCCTTTGAATCAAAAGGATTAGATGCCTTTGTCGCTGGT
ATCGGTACAGGCGGCACGATCACTGGTGCGGGTCGAGAATTAAAACGAGTGTATCCAGACTTAAAAATCTATGGCGTTGAACCGGCAGAATCGGCTATCTTAGAAGGCGGACAACCTGGTCCTCACAA
AATCCAAGGAATTGGAACCGGCTTTGTACCTAAAGTTCTTGATACAGAGATCTTTGATCAAGCAATTGCCGTATCAAGTGAAGATGCGATGGCAACAGCGAGAGAAGTCGGCCGTAAAGAGGGCTTGC
TCGTAGGTATTTCTTCTGGAGCAGCGATCCATGCAGCACTCAAAGTTGCCAAAGAACTGGGAAAAGGCAAAAAAGTCTTAGCGATCGTCCCGGACAATGGAGAACGCTATCTATCGACTGCTTTATAT
CAAGTGGAAGAATAACAAGGAAAGCGGGCATTGGCAGAATGCCTGCTTTTTCTCCCCTTTTCTTAAGGACTGTGGAATTTTTCGTGTTTGTTTTGTGGTTGATTAAGAATTTTTAGAGAAAGCTGTTT
GAATTTTTACCATCTACATAGTATGATATTACAGATAAGAGAATTGGAGGAACTCCTGTGAACACGATATATGATGAGGCAATCAAACAATTAAAACAAGAAAATATCCGTATCACTCCACAAAGGGT
TGCGATACTTGAGTTTTTGGCTAGCCACGCATCACATCCTACCGCAGAAGAAATTTACCGAGCGATCGAAGTCCACTTTCCCGGTATTAGTGTAGCGACAGTCTACAATAATCTGCGACTGTTTACGG
AGATCGGTTTTGTGAAAGAAATGAACTACGGCGATGCATCCAGTCGCTTTGACTTTAAGACAGAGCCGCATTATCATGCAATTTGTCAACGTTGCGGCAAGATTTCTGATTTTTACTATCCTGGATTG
GAAGATGTCGAAATGGCCGCTTCCCAACTGACCGGATATAAAATCCAAAGCCATCGCTTAGAAGTATATGGATTGTGTCCAGAGTGTAAAGGAGAAGAGGAATGAAAATTATTTCAACGGATAAAGCA
CCGCAAGCAATTGGTCCTTATGTACAAGCACGGATTGCTAATGGTTTCTTGTTTGCTTCCGGACAAGTGCCGTTAGATCCTGTTACAGGAGAAGTTGTGGGAGCAACTATCGAAGTTCAGACACAGCA
AGTTCTCAAAAATATCGAGGCAATTTTAACTGCCGCCGGTATTACACCAGATGAAATCGTCAAAACAACATGTTTCTTAAAAAACATGGATGATTTTGTTCCATTCAATAATACATACAGCACACTTT
TTCATGAACATTTGCCTGCACGTTCAGCAGTTGAAGTGGCCCGTTTACCAAAGGATGTATTGGTGGAAGTAGAGATCATTGCTGTGGTGGCACAATAAAGGAGTCAATGGGACGGTCGTATATATGCG
CAGTCCCATTTTTGTTGAGAATGGATTTTATTCTTGATGATGCCAAAATAAAATTGAAAATGGTTTTTATGTAACGATGGGAAAAGAGGACAATAAGATGATTTCTTTACAAGAATATTTAGCCCAGG
GGCAAACAACTGTCTCCAATCTGCTTTTGGAAAACTATCATAGCCTTGGTTTAACCAATGATGAATTTGTATTATGGCTTCAGTTGTATCGTTACCACGAACAAGGCAATGATTTTCCAGATTTAGCG
ATGATTGCTAAAGGGATGGGAAGTGAGCAAAAAGAAATCTATAAATGGTTGAATGAGCTATTGAATAAACAGATCATCGCCATCGAATCCAAGAAAGATTCAGCAGGGAAGATGGTGGATTACTATGA
GTTTTCTGGAATCTATGAACGACTGGAAGTTTGTTTGCGGCAAAAAAATCAGCAGGAGGAAAAGCAGAGCTATGATTCAAAAGTACGTTCGCTTTATCAAATGTTTGAGTCCGAATTTGGCCGTCCGC
TTTCAGCAATCGAATACCAGCGAATATCACAATGGCTGGAAGAAGATAAATACCAGCCGGATTTGATTCAATTTGCGTTACGTGAGGCGGTTTTGAATCAAGCCTACAGTTTGAATTATATTGATCGC
ATTTTACTTTCTTGGGAACGGAAAAATATTACCTCGAAAGAACAAGTAGAAGAAGAACAAAAACGCCGTAAAAAACAGTTTATGCAGAAAGAAAATTTGGAACAAACGAATGGACCGGTTCCTAAGAT
TCCTATGCATAATTGGTTAGAAGGGGAATAAAGGAGAGTTGCCATGATCAATAAAGCAAAAACTATGATCGCATTGGAACAAATGTACCAAATGTTCCCCGACGCTCACGGAGAACTGATTTCCAAAA
ATCCGTTTGAACTGTTGATTGCTGTGATTTTAAGTGCTCAGGCTACGGATGTGTCGGTCAATAAAGTGACCCCCACACTATTTGCTGCCTATCCGACACCTGAGGCCTTGGCAGCAGCTCCTGTTGAG
GAAATCATTGAGAAAATACGGACGATCGGTCTTTATCGCAACAAAGCAAAAAATATCAAGGCTTGTGCGTCCCAGCTTATCGAACGATTCAATGGCCAAGTACCTAGAACCCGCGAGGAATTAGTGTC
ATTACCGGGGGTCGGAAGAAAAACAGCCAACGTTGTTTTGGGCGATGCGTTTGGTATTCCGGCAATCGCTGTGGATACCCACGTCGAACGAGTAACCAAGCGCTTACGGATTTGCCGATTAGATGCAA
ATGTCTTAGAAGTAGAACAAACATTGATGAAAAAAGTGCCAGAAGATCTTTGGGTATTGATGAAAAAAGTGCCAGAAGATCTTTGGGTCAAAACACATCACACACTGATTTTCTTTGGACGATACCAT
TGCACAGCTAGAGCGCCAAAATGTGAAGTTTGCCCATTATTGACAATGTGCCAAGAAGGACAACTGCGGATGCGCGCAAAAAAAGTTCAGCCACCTCAATGA

#Corrected sequence

>L1_enterococcussp.hsieg1_1_EQC80769.1_upstream_e_DNA_coding_sequence

ATGATCAATAAAGCAAAAACTATGATCGCATTGGAACAAATGTACCAAATGTTCCCCGACGCTCACGGAGAACTGATTTCCAAAAATCCGTTTGAACTGTTGATTGCTGTGATTTTAAGTGCTCAGGC
TACGGATGTGTCGGTCAATAAAGTGACCCCCACACTATTTGCTGCCTATCCGACACCTGAGGCCTTGGCAGCAGCTCCTGTTGAGGAAATCATTGAGAAAATACGGACGATCGGTCTTTATCGCAACA
AAGCAAAAAATATCAAGGCTTGTGCGTCCCAGCTTATCGAACGATTCAATGGCCAAGTACCTAGAACCCGCGAGGAATTAGTGTCATTACCGGGGGGTCGGAAGAAAAACAGCCAACGTTGTTTTGGGC
GATGCGTTTGGTATTCCGGCAATCGCTGTGGATACCCACGTCGAACGAGTAACCAAGCGCTTACGGATTTGCCGATTAGATGCAAATGTCTTAGAAGTAGAACAAACATTGATGAAAAAAGTGCCAGA
AGATCTTTGGGTATTGATGAAAAAAGTGCCAGAAGATCTTTGGGTCAAAACACATCACACACTGATTTTCTTTGGACGATACCATTGCACAGCTAGAGCGCCAAAATGTGAAGTTTGCCCATTATTGA
CAATGTGCCAAGAAGGACAACTGCGGATGCGCGCAAAAAAAGTTCAGCCACCTCAATGA

>L1_enterococcussp.hsieg1_1_EQC80769.1_protein_sequence_

MINKAKTMIALEQMYQMFPDAHGELISKNPFELLIAVILSAQATDVSVNKVTPTLFAAYP
TPEALAAAPVEEIIEKIRTIGLYRNKAKNIKACASQLIERFNGQVPRTREELVSLPGVGR
KTANVVLGDAFGIPAIAVDTHVERVTKRLRICRLDANVLEVEQTLMKKVPEDLWVLMKKV
PEDLWVKTHHTLIFFGRYHCTARAPKCEVCPLLTMCQEGQLRMRAKKVQPPQ*

>L1_enterococcussp.fdaargos_553_1_AYY10705.1_protein_sequence

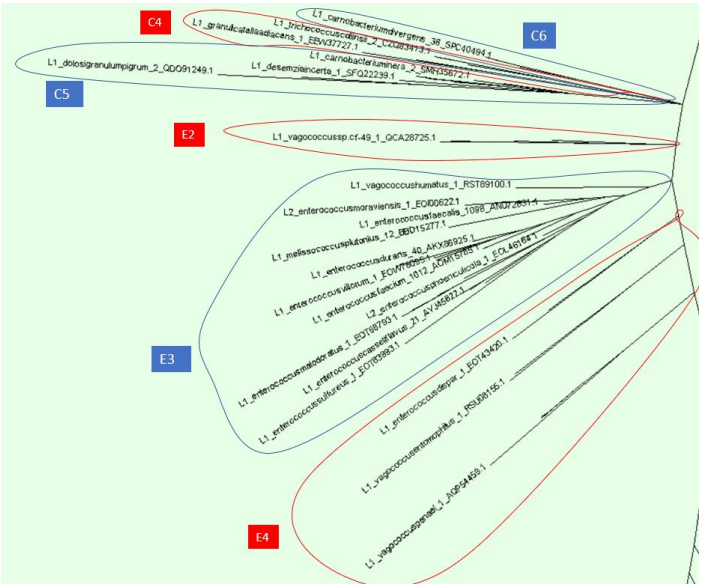
MISKAKTMIALEQMYQMFPDAHGELISKNPFELLIAVILSAQATDVSVNKVTPTLFAAYPTPEALAAAPVEEIIEKIRTIGLYRNKAKNIKACASQLIERFNGQVPRTREELVSLPGVGRKTANVVLG
DAFGIPAIAVDTHVERVTKRLRICRLDANVLEVEQTLMKKVPEDLWVKTHHTLIFFGRYHCTARAPKCEVCPLLTMCQEGQLRMRAKKVQPPQ*

```
Output file   outfile

    ########################################
    # Program: needle
    # Rundate: Mon 18 Apr 2022 15:15:58
    # Commandline: needle
    #    -auto
    #    -asequence /var/lib/emboss-explorer/output/778444/.asequence
    #    -bsequence /var/lib/emboss-explorer/output/778444/.bsequence
    #    -nocendweight
    #    -brief
    #    -outfile outfile
    #    -aformat3 srspair
    # Align_format: srspair
    # Report_file: outfile
    ########################################

    #=======================================
    #
    # Aligned_sequences: 2
    # 1: L1_enterococcussp.hsieg1_1_EQC80769.1_upstream_e_DNA_coding_sequence_1
    # 2: L1_enterococcussp.fdaargos_553_1_AYY10705.1_protein_sequence
    # Matrix: EBLOSUM62
    # Gap_penalty: 10.0
    # Extend_penalty: 0.5
    #
    # Length: 233
    # Identity:     221/233 (94.8%)
    # Similarity:   222/233 (95.3%)
    # Gaps:          11/233 ( 4.7%)
    # Score: 1114.0
    #
    #
    #=======================================

    L1_enterococc      1 MINKAKTMIALEQMYQMFPDAHGELISKNPFELLIAVILSAQATDVSVNK     50
                         ||:||||||||||||||||||||||||||||||||||||||||||||||||
    L1_enterococc      1 MISKAKTMIALEQMYQMFPDAHGELISKNPFELLIAVILSAQATDVSVNK     50

    L1_enterococc     51 VTPTLFAAYPTPEALAAAPVEEIIEKIRTIGLYRNKAKNIKACASQLIER    100
                         ||||||||||||||||||||||||||||||||||||||||||||||||||||
    L1_enterococc     51 VTPTLFAAYPTPEALAAAPVEEIIEKIRTIGLYRNKAKNIKACASQLIER    100

    L1_enterococc    101 FNGQVPRTREELVSLPGVGRKTANVVLGDAFGIPAIAVDTHVERVTKRLR    150
                         ||||||||||||||||||||||||||||||||||||||||||||||||||||
    L1_enterococc    101 FNGQVPRTREELVSLPGVGRKTANVVLGDAFGIPAIAVDTHVERVTKRLR    150

    L1_enterococc    151 ICRLDANVLEVEQTLMKKVPEDLWVLMKKVPEDLWVKTHHTLIFFGRYHC    200
                         |||||||||||||           |||||||||||||||||||||||||
    L1_enterococc    151 ICRLDANVLEVEQT-----------LMKKVPEDLWVKTHHTLIFFGRYHC    189

    L1_enterococc    201 TARAPKCEVCPLLTMCQEGQLRMRAKKVQPPQ*    233
                         ||||||||||||||||||||||||||||||||
    L1_enterococc    190 TARAPKCEVCPLLTMCQEGQLRMRAKKVQPPQ*    222
```

Annex 3- *Enterococcus sp. HSIEG1 HSIEG1* EQC80769.1 sequence correction with a close gene, *Enterococcus sp. FDAARGOS_375 FDAARGOS_375* _AYY10705.1 obtained by Emboss Needle and Transeq.

```
L1_enterococc      1 ------MISKAKTMIALEQMYQMFPDAHGELISKNPFELLIAVILSAQAT     44
                           .:.:..|:...:::.:. |.||:...|:..:|..:.:..
L1_streptococ      1 MHAVYGNVLEGLTLKQCQEIHAVI-DEHGKAFGKSKGQLDESWLR*----     45

L1_enterococc     45 DVSVNKVTPTLFAAYPTPEALAAAPVEEIIEKIRTIGLYRNKAKNIKACA     94

L1_streptococ     45 -------------------------------------------------     45

L1_enterococc     95 SQLIERFNGQVPRTREELVSLPGVGRKTANVVLGDAFGIPAIAVDTHVER    144

L1_streptococ     45 -------------------------------------------------     45

L1_enterococc    145 VTKRLRICRLDANVLEVEQTLMKKVPEDLWVKTHHTLIFFGRYHCTARAP    194

L1_streptococ     45 -------------------------------------------------     45

L1_enterococc    195 KCEVCPLLTMCQEGQLRMRAKKVQPPQ*    222

L1_streptococ     45 ----------------------------     45
```

Annex 4- *Streptococcus sp. HSISM1 HSISM1* EQC74190.1 sequence alignment with the reference, *Enterococcus sp. FDAARGOS_553 FDAARGOS_553* AYY10705.1. The lines represent common amino acids, and the dots represent lack of alignment due to different amino acids in comparison.



Annex 5-Middle part of the phylogenetic tree from Figure 6 showing in detail the composition the clusters C4, C5, C6 and E2, E3 and E4 represented.

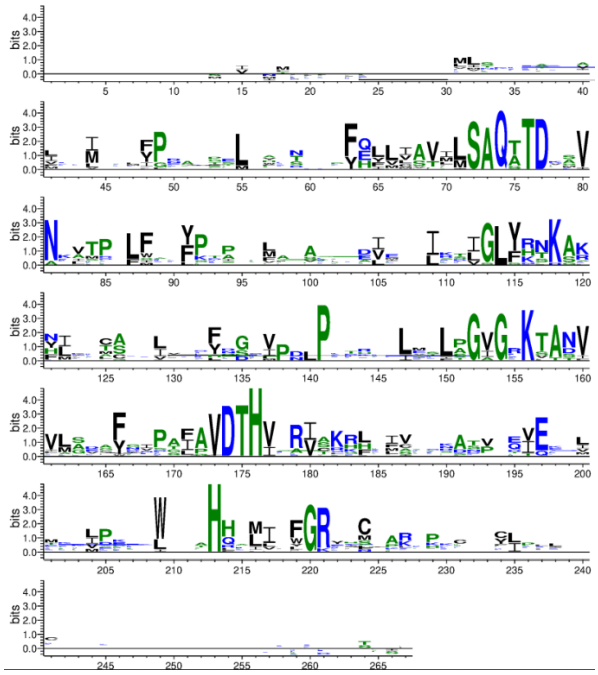
		Clusters																	
Taxonomic Family	species and strain	A1	A2	C1	C3	C4	C5	C6	CA1	E2	E3	E4	EA1	L1	L2	L3	S1	S2	SOMA
Aerococcacea	L1_aerococcaceaebacterium_2	1											1						2
	L1_aerococcuschristensenii_1	2																	2
	L1_aerococcusurinae_1	2																	2
	L1_aerococcussanguinicola_1	2																	2
	L1_aerococcusurinae_4	2																	2
	L1_aerococcusurinaehominis_1	2																	2
	L1_facklamiahominis_2	1																	1
	L1_facklamiaignava_1	1																	1
	L1_facklamialanguida_1	1																	1
	L1_abiotrophiadefectiva_1		1																1
	L1_aerococcusurinaeequi_2								1										1
	L1_aerococcusurinaeequi_5								1										1
	L1_aerococcusviridans_3								1										1
	L1_aerococcussuis_1												1						1

		Clusters																	
Taxonomic Family	species and strain	A1	A2	C1	C3	C4	C5	C6	CA1	E2	E3	E4	EA1	L1	L2	L3	S1	S2	SOMA
Carnobacteriaceae	L1_jeotgalibacadankookensis_2			1															1
	L1_jeotgalibacasp.h21t32_1			1															1
	L1_jeotgalibacasp.pts2502_1			1															1
	L1_carnobacteriumalterfunditum_1				1		1												2
	L1_carnobacteriuminers_2				1		1		1										3
	L1_carnobacteriuminhibens_1				2		1												3
	L1_carnobacteriumsp.17-4_1				2		1												3
	L1_carnobacteriumsp.cp1_1				1		1												2
	L1_carnobacteriumviridans_1				1		1												2
	L1_desemziaincerta_1				1		1												2
	L1_granulicatellaadiacens_1					1													1
	L1_granulicatellaelegans_1					1													1
	L1_trichococcuscollinsii_2					1													1
	L1_trichococcuspalustris_2					1													1
	L1_trichococcussp.skbg_1					1													1
	L1_dolosigranulumpigrum_2						1												1
	L1_marinilactibacillussp.15r_1						1												1
	L1_carnobacteriumdivergens_38							1											1
	L1_carnobacteriumdivergens_41							1											1
	L1_carnobacteriummaltaromaticum_17							1											1
	L1_carnobacteriummaltaromaticum_30							1											1
	L1_carnobacteriumsp.pl12red10_1								1										1
	L1_carnobacteriumsp.pl17gre32_1								1										1
	L1_carnobacteriumsp.pl17red31_1								1										1
	L1_carnobacteriumsp.pl24red07_1								1										1
	L1_carnobacteriumsp.pl26red25_1								1										1

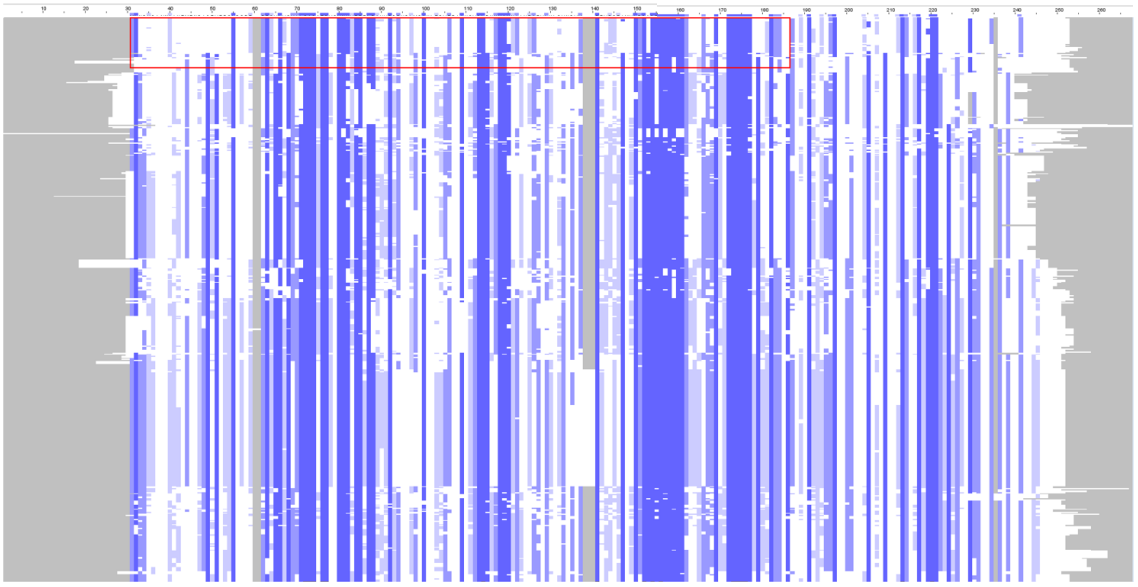
		Clusters																	
Taxonomic Family	species and strain	A1	A2	C1	C3	C4	C5	C6	CA1	E2	E3	E4	EA1	L1	L2	L3	S1	S2	SOMA
	L1_vagococcusfessus_1									1									1
	L1_vagococcuslutrae_2									1									1
	L1_vagococcussp.cf-49_1									1									1
	L1_vagococcussp.mn-17_1									1									1
	L1_enterococcusavium_1										1								1
	L1_enterococcusavium_13										1								1
	L1_enterococcuscasseliflavus_21										1								1
	L1_enterococcuscasseliflavus_23										1								1
	L1_enterococcusdurans_22										1								1
	L1_enterococcusdurans_39										1								1
	L1_enterococcusdurans_40										1								1
	L1_enterococcusdurans_48										1								1
	L1_enterococcusfaecalis_1027										1								1
	L1_enterococcusfaecalis_1060										1								1
	L1_enterococcusfaecalis_1065										1								1
	L1_enterococcusfaecalis_1098										1								1
	L1_enterococcusfaecalis_1346										1								1
	L1_enterococcusfaecalis_659										1								1
	L1_enterococcusfaecalis_882										1								1
	L1_enterococcusfaecium_1012										1								1
	L1_enterococcusfaecium_127										1								1
	L1_enterococcusfaecium_128										1								1
	L1_enterococcusfaecium_389										1								1
	L1_enterococcusfaecium_43										1								1
	L1_enterococcusfaecium_524										1								1
	L1_enterococcusfaecium_535										1								1
	L1_enterococcusgallinarum_18										1								1
	L1_enterococcusgilvus_2										1								1
	L1_enterococcushaemoperoxidus_1										1								1
	L1_enterococcushirae_27										1								1
	L1_enterococcushirae_41										1								1
	L1_enterococcushirae_50										1								1
	L1_enterococcushirae_59										1								1
Enterococcaceae	L1_enterococcusmalodoratus_1										1								1
	L1_enterococcusmundtii_12										1								1
	L1_enterococcusmundtii_21										1								1
	L1_enterococcusmundtii_22										1								1
	L1_enterococcusmundtii_24										1								1
	L1_enterococcusquebecensis_2										1								1
	L1_enterococcusrotai_1										1								1
	L1_enterococcussaccharolyticus_4										1								1
	L1_enterococcussilesiacus_2										1								1
	L1_enterococcussp.cr-ec1_1										1								1
	L1_enterococcussp.fdaargos_375										1								1
	L1_enterococcussp.fdaargos_553										1								1
	L1_enterococcussp.hsieg1_1										1								1
	L1_enterococcussp.m190262_1										1								1
	L1_enterococcussulfureus_1										1								1
	L1_enterococcusthailandicus_1										1								1
	L1_enterococcusureasiticus_1										1								1
	L1_enterococcusvillorum_1										1								1
	L1_melissococcusplutonius_12										1								1
	L1_melissococcusplutonius_16										1								1
	L1_melissococcusplutonius_9										1								1
	L1_vagococcushumatus_1										1								1
	L2_enterococcusfaecalis_1386										1								1
	L2_enterococcusmoraviensis_1										1								1
	L2_enterococcusphoeniculicola_1										1								1
	L2_enterococcusraffinosus_1										1								1
	L1_enterococcusasini_2											1							1
	L1_enterococcuscecorum_21											1							1
	L1_enterococcusdispar_1											1							1
	L1_melissococcussp.om08-11bh_1											1							1
	L1_vagococcusentomophilus_1											1							1
	L1_vagococcusmartis_1											1							1
	L1_vagococcuspenaei_1											1							1
	L1_vagococcusteuberi_1											1							1
	L1_tetragenococcushalophilus_5												1						1
	L1_tetragenococcuskoreensis_1												1						1
	L1_tetragenococcusosmophilus_1												1						1

		Clusters																	
Taxonomic Family	species and strain	A1	A2	C1	C3	C4	C5	C6	CA1	E2	E3	E4	EA1	L1	L2	L3	S1	S2	SOMA
	L1_lactobacillusbrantae_1													1					1
	L1_lactobacilluscasei_1													1					1
	L1_lactobacilluscasei_11													1					1
	L1_lactobacilluscasei_19													1					1
	L1_lactobacilluscasei_25													1					1
	L1_lactobacilluscasei_6													1					1
	L1_lactobacillusmanihotivorans_3													1					1
Lactococcaceae	L1_lactobacillusparacasei_101													1					1
	L1_lactobacillusparacasei_107													1					1
	L1_lactobacillusparacasei_11													1					1
	L1_lactobacillusparacasei_126													1					1
	L1_lactobacillusparacasei_78													1					1
	L1_leuconostocmesenteroides_14													1					1
	L1_lactobacillusrhamnosus_48													1					1
	L1_lactobacillusrhamnosus_71													1					1

Annex 6- Composition of the Nth1 phylogenetic tree divided into clusters represented in detail.



Annex 7– Logos of the Nth1, the area occupied by each letter indicates the background frequency of an amino acid in the input sequence set.



Annex 8- Overview of the multiple alignment, generated by Muscle, of the Nth1 identified in this research. The intensity of the colour blue represents the frequency of the sequence as does the size of the amino acid represented in the logos.