Comparative genomic study reveals the signature collinsella regions coding for sugar-degrading enzymes and key differences within and outside the genus

George Alehandro Saad [a*]
[a] Faculty of Sciences, Aix-Marseille University, Marseille, France
* Email address: george-alehandro.saad@etu.univ-amu.fr

A. Abstract

Comparative genomics is a branch of genomics that involves examining the genomes of several organisms to identify genetic similarities and differences as well as their evolutionary ties. This can be achieved by investigating the DNA composition, gene content, and genome organization of various organisms. In this report, we performed a comparative genomics analysis of Colinsella aerofaciens and other closely related species in the genus Collinsella and other species that belong to the same family but to a different genus. Our analysis revealed significant genomic differences among these bacteria, including differences in gene content, gene expression, and metabolic pathways. Our results suggest that these genomic differences may contribute to the functional diversity of Collinsella species and their potential roles in human health and disease. Our study highlights the potential of comparative genomics as a tool for understanding the evolution and functional potential of Collinsella species and other bacteria in the human gut microbiome.

Keywords: comparative genomics, collinsella, senegalimassilia, coriobacteriaceae, orthologous

B. INTRODUCTION

Collinsella aerofaciens belongs to the genus Collinsella within the family Coriobacteriaceae. The purpose of this report is to shed the light on the characteristics of this bacteria and its genus by using comparative genomics methods[1]. To do so, two species belonging to the same genus (Collinsella provencensis[2] and Collinsella intestinalis[3]) and another two belonging to the same family but to different genus (Senegalimassilia anaerobia[4], Senegalimassilia faecalis[5]) were selected. These species belong to the gut microbiome. Many methods of comparative genomics were applied between these species to further understand the features of Collinsella aerofaciens. C. aerofaciens is one of the genomes that are available in the COG database[6].

1. Description

Collinsella is a genus of Gram-positive, non-spore-forming, obligately anaerobic bacteria in the family Coriobacteriaceae. These bacteria are often found in the human gut microbiome and have been identified in the feces of healthy individuals. They are also found in soil, water, and the feces of animals.

The genus Collinsella is an important component of the human gut microbiome, and it is thought to play a role in maintaining gut health and influencing the host immune system.

These species are rod-shaped and typically measure between 0.5 and 0.7 micrometers in width and between 1.5 and 5.0 micrometers in length. They are non-motile and do not produce endospores. They are obligately anaerobic, meaning they require an environment that is free of oxygen in order to grow and thrive.

They are known to produce a range of enzymes that allow them to break down complex sugars and other compounds that are found in the human gut. This metabolic activity is thought to contribute to the overall health of the human gut microbiome.

Collinsella species are typically facultative anaerobes, meaning they can grow in the presence or absence of oxygen. They are often found in mixed communities with other bacteria, and they are known to produce enzymes that help them break down complex carbohydrates and other biomolecules.

2. Pathogenicity

Collinsella species are generally considered to be non-pathogenic, meaning they do not cause disease in humans or animals. However, some species have been associated with infections in immunocompromised individuals, such as those with HIV/AIDS or cancer[7].

It's also worth mentioning that some species of Collinsella have been found to produce virulence factors, which are molecules or substances that help the bacterium establish an infection and evade the host immune system. However, the role of these virulence factors in the pathogenicity of Collinsella species is not well understood, and more research is needed to fully understand the potential pathogenic mechanisms of these bacteria.

3. History

The earliest description of Collinsella Aerofaciens was made by Eggerth 1935 [8] as “Eubacterium Aerofaciens”. It was later transferred to the genus of Collinsella as Collinsella aerofaciens[9] in 1999.The C11 strain that will be analyzed in our study was isolated in Seoul, South Korea in July 2007 by the "Korea Food Research Institute”.

C. MATERIALS AND METHODS

1. Materials

1.1. Data

To study the genome of a certain species, the data is retrieved from the NCBI’s refseq database[10]. The collection of the files are as following (table 1).

File ending with	Description
_assembly_report.txt	Reports the assembly status of this version of the assembly
_assembly_stats.txt	Contains general information about the species and specific stats (total length…)
_cds_from_genomic.fna	Nucleotide FASTA of all the annotated CDS features
feature table.txt	Describes the features (Gene, CDS…) of the genome with information like their strand, start, end and GeneID...
_feature_count.txt	Based on feature_table.txt will generate descriptive counting statistics
_genomic.fna	Nucleotide FASTA of the genome in the assembly
_genomic.gbff	GenBank file format
_genomic.gff	Annotation of the genome in GFF3
_genomic.gtf	Annotation of the genome in GTF2.2
_protein.faa	Amino Acid FASTA of all the protein products of the genome
_protein.gpff	GenPept format of the protein products
_rna_from_genomic.fna	Nucleotide FASTA of all the RNA products annotated
_translated_cds.faa	Amino Acid FASTA of all the CDS sequences of the genome
md5checksums.txt	checksums used to verify data integrity
README.md	Markdown description of the files and their columns
README.txt	Description of the files and their columns

Table 1: Description of the files retrieved from the NCBI refseq database

The retrieved data of the five organisms is extracted from the NCBI database. The name of the species, strains, refesq assembly and link to the ftp repositories are shown in table 2.

	Bacteria	Strain	RefSeq Assembly	Link to data resources
Genome of interest	Collinsella aerofaciens	C11	GCF_003856815.1	Link
Same genus	Collinsella provencensis	Marseille-P3740	GCF_900199705.1	Link
	Collinsella intenstinalis	DSM-13632	GCF_902501455.1	Link
Same family but different genus	Senegalimassilia anaerobia	JC110	GCF_000236865.1	Link
	Senegalimassilia faecalis	KGMB04484	GCF_004135645.1	Link

Table 2: General information about the species in this study with the link to the data

After accessing the links of these genomes, the files can be retrieved either manually one by one or by software like FileZilla which allows the connection to ftp servers and the download of all the files in one click.

1.2. Software Analysis

For the purpose of fetching, generating statistics and exploring the various relations between the different species that are being analyzed in our study, multiple Software (table 3), either based on a web server or downloaded locally, were used for that.

Tool	Objective	Data input	Data output	Parameters
NCBI	Retrieves the genomes' data	Name of the organism query	Folder containing all the files of the genome	-
Compseq (EMBOSS)[11]	Calculates the ratio between the frequencies of di-nucleotides in the genome compared to the randomly expected frequencies	*_genomic.fna	Txt file	Default
D-Genies[12]	Aligning and compare multiple DNA sequences, identify conserved regions and structural features,	*_genomic.fna	.png plot .paf file containing coordinates	
Orthofinder[13]	Inferring orthogroups of protein coding genes and constructing a gene tree	*_protein.faa	Directory containing various files formats	-
eggNOG-mapper 2.19 [14]	Infers a COG classification for the proteins of the studied species	* _protein.faa	Folder containing multiple files indicating the annotations inferred	Many to many Default 0.001 Min e-hit value 60 Min hit bit-score 40 Min % of identity 20 Min % of query coverage 20 Min % of subject coverage
Database of Clusters of Orthologous Genes (COGs) NCBI[6]	Fetching the orthologous group classification of Collinsella Aerofaciens proteins based on COGs.	Collinsella Aerofaciens	Cog_result_table.tsv	-
Orthovenn2 [15]	Clustering technique for Identifying orthologous gene clusters Can determine core, essential and specific gene groups for all the 5 species	5 proteomes : *_translated_CDS.faa	Plot	Default
Cusp (EMBOSS)[11]	Calculates the RSCU	*_cds_from_genomic.fna	Txt file	-
CAI (EMBOSS)[11]	Calculates the CAI	*_cds_from_genomic.fna	Txt file	-

Table 3: List of the tools used with each of their objectives, input, output and parameters

1.3. Other Software

Beside the web-based tools used for the genomic comparative approaches, other tools were used locally for data exploration and more specific tasks as shown in table 4. For detailed info about the usage of Python (Annex 4) and Excel (Annex 5) please refer to the annex and additional files.

Tool	Objective	Functions
Excel 2013	Organize tables Generate Statistics	-
Python 3.9	Data mining	extract_stats
	Length and %GC	length_and_gc
	Extracting the number of COGs categories retrieved by EggNOG	cog_categories_EGGNOG
	Comparing COG annotation of Collinsella Aerofaciens VS the annotation inferred by EggNOG	cog_categories
	Statistics around the CDS	CDS_coding_protein_stats
	Statistics around the genes	gene_stats
	Extracting best hits for BlastP results after filtering (% identity > 30 and alignment length > %80 of the smaller sequence) Directly calculates the AAI	BlastpAlignmentLengthFilter
	Extracting best hits for BlastN results after filtering (% identity >= 70 and alignment length >= %70 of the smaller sequence)	BlastnAlignmentLengthFilter
	Calculating the ANI between the genome of interest and all the other genomes one by one	CalculateANI
	Calculating the ANI between the genome of interest and all the other genomes one by one	Calculate AF

Table 4: Description of the usage of Excel and Python in our report

1.4. Operating System

The local analysis was done in two different operating systems: Windows 10 and Ubuntu 20.10.
The reason being that some of the software installed on local were easier to install on Ubuntu, coupled with the usage of bash for manipulating the files. Some of the tools that were installed on local Ubuntu: NCBI's blast and Orthofinder. While on Windows it was mainly easier to use Excel, generate plots and launch analysis on web-based tools.

2. Methods

2.1. Comparative Analysis of genome structure

2.1.1. General statistics and description

A descriptive approach for studying the characteristics of each of the analyzed species. This method will be based on the downloaded files from NCBI and it will showcase diverse statistics surrounding the genomic sequence, the proteins, the RNAs, the genes and pseudogenes of these species.

2.1.2. Frequency of usage of dinucleotides

Studying the frequency of dinucleotide usage in comparative genomics can provide insights into the evolutionary history and molecular biology of different organisms. We will use Compseq from EMBOSS[11]. Dinucleotide frequency, also known as dinucleotide composition, is a measure of the occurrence of pairs of nucleotides within a DNA sequence. It can be used to compare the DNA sequences of different organisms to identify patterns and trends in their evolution. It can be used to identify differences between closely related organisms, such as between different strains or species, and can provide insights into the mechanisms of evolution and speciation. The method used here compares the frequency of the dinucleotide observed divided by the frequency that this nucleotide is observed in an equal theoretical proportion.

$$value = \frac{frequency\ observed}{frequency\ expected}$$

With frequency expected = $\frac{4*4}{2\text{^}4*2\text{^}4}$ = 0.0625 (Equal probability of obtaining one of the four nucleotides A, T, C, G).

2.1.3. Sequence alignment of genomes

Aligning two genome sequences can also help in identifying structural differences between the genomes, such as insertions, deletions, or rearrangements. These differences may be important for understanding how the genomes of different organisms have evolved over time. D-GENIES[12] is a software tool that is used for analyzing and comparing DNA sequences. It is designed to be a user-friendly platform for researchers and biologists to perform various types of analysis on DNA sequences, including alignment, annotation, and phylogenetic tree construction. D-GENIES is user-friendly and easy to use, with a graphical user interface (GUI) that allows researchers to visualize and interact with the results of their analyses. After aligning with D-GENIES, the resulting aligned sequences can be retrieved in .paf format. The coordinates of the aligned sequences are retrieved and then by using the feature table of the genome of interest (C. aerofaciens), it's possible to extract a functional information about the role of these sequences.

2.1.4. BlastN

BlastN is a program within the BLAST suite that is used to search for similarity between nucleotide sequences. It works by taking a query nucleotide sequence and comparing it to a database of nucleotide sequences, looking for regions of similarity between the two. The program returns a list of sequences from the database that match the query, along with information about the degree of similarity and the location of the matching regions. It will be used implicitly during the calculation of ANI and AF.

2.2. Comparative Analysis of the proteome

2.2.1. Retrieving the OG annotation of all the proteomes

The EggNOG-mapper tool [14] takes a set of protein sequences as input and uses them to search the EggNOG database for matches to known orthologous genes. It then assigns these genes to functional categories (called "Orthologous Groups" or "OGs") based on their evolutionary relationships and known functions. The tool also provides annotations and other information about the genes, such as their predicted function, taxonomic distribution, and evolutionary history. This step will allows the comparison between the frequencies of the classifications between the studied species.

2.2.2. Comparing the COG annotation with the EggNOG-mapper

Since the C. aerofaciens specie is completely annotated in the NCBI's COG database, it was interesting to make a comparison between the COG's annotation and the EggNOG-mapper inferred annotation. This method will manifest the differences between the manually curated resource annotation of COG and the automatic classifications by EggNOG-mapper. It's possible to compare quantitatively and qualitatively of both approaches. Also one additional metric can be observed: the annotation rate. The annotation rate can be calculated as following:

$$annotation\ rate = \frac{number\ of\ proteins\ annotated}{total\ number\ of\ proteins\ of\ the\ proteome}$$

It can be useful to compare the abundancy of proteins that have been annotated.

2.2.3. Comparing orthologous groups across multiple genomes

OrthoVenn2[15] is particularly useful for comparative genomics studies, as it allows researchers to identify and compare the genes that are present in different organisms and to gain insights into the evolutionary relationships between the organisms. It is also useful for functional genomics studies, as it can help in identifying and classifying genes based on their functions and evolutionary relationships. We input a list of orthologous groups and the genomes that we are interested in comparing. It then generates a graphical representation of the orthologous groups, showing which genes are present in each genome and how the genes are distributed among the genomes. This allows us to easily visualize and compare the content of the orthologous groups across the genomes and to identify patterns and trends in the data.

2.2.4. Finding Orthologous groups with gene and phylogenetic tree inference

OrthoFinder[13], [16] uses sequence similarity and functional annotations to predict orthologs and can be used to study the evolution of gene families, investigate the functions of newly sequenced genomes, and predict gene functions in uncharacterized genomes. It also infers a rooted species tree for the species being analyzed and maps the gene duplication events from the gene trees to branches in the species tree. OrthoFinder performs a multiple

sequence alignment (MSA) of the protein sequences and uses the Markov Cluster Algorithm (MCL) to group the sequences into clusters based on their similarity. These clusters can then be used to infer gene trees and rooted species trees by reconstructing the evolutionary relationships between the genes.
To get the Resolved Gene Trees OrthoFinder carries out a Duplication-Loss-Coalescence analysis to identify the more parsimonious interpretation of the tree. Experimentally, in order

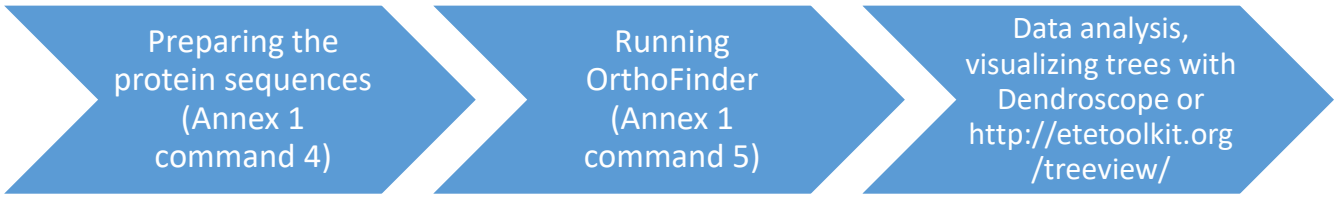


Figure 1: The steps of running an Orthofinder analysis (commands in annex)

2.2.5. BlastP

BlastP is used to help identify the function of a protein or to study evolutionary relationships between different species. It is a powerful tool for analyzing large amounts of protein data and can be used to identify proteins with similar functions, predict protein-protein interactions, and study evolutionary relationships between organisms. It will be used implicitly during the calculation of AAI.

2.3. Comparing the coding sequences

2.3.1. Calculating ANI

Average nucleotide identity (ANI) is a measure of the similarity between two DNA sequences. It is calculated by aligning the sequences and comparing the nucleotides at each position, by BlastN, and is typically expressed as a percentage. In our approach, the best hit for each sequence of the genome pair is taken having %70 or more identity and at least %70 coverage of the shorter gene. A higher ANI value indicates a higher level of sequence similarity between the two sequences, while a lower ANI value indicates a lower level of sequence similarity.

$$ANI = \frac{\sum(Percent\ identity * alignment\ length)}{length\ of\ BH\ gene}$$

The steps of calculating the ANI are shown in figure 2 below.

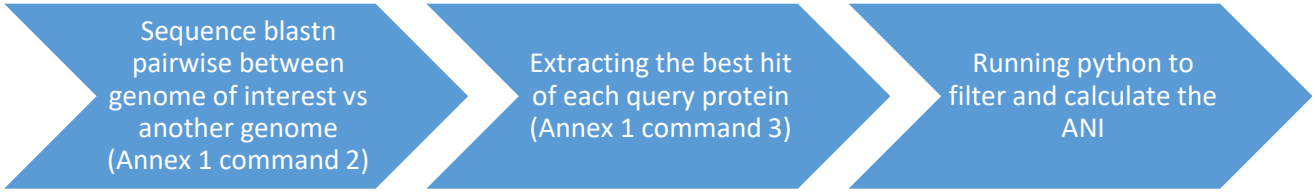


Figure 2: Steps of calculating ANI (commands in annex)

2.3.2. Calculating AAI

Average amino acid identity (AAI) is a measure of the similarity between two protein sequences. The sequences are compared by BlastP and with a threshold of more than %30 identity and coverage of more than %80 of the shorter gene the sequences are then filtered to extract only the best hit (BH). Also expressed by a percentage, a higher AAI value indicates a higher level of sequence similarity between the two proteins, while a lower AAI value indicates a lower level of sequence similarity.

$$AAI = \frac{\sum(Percent\ identity * alignment\ length)}{length\ of\ BH\ gene}$$

Similarly to ANI, there are some steps that are done (figure 3).

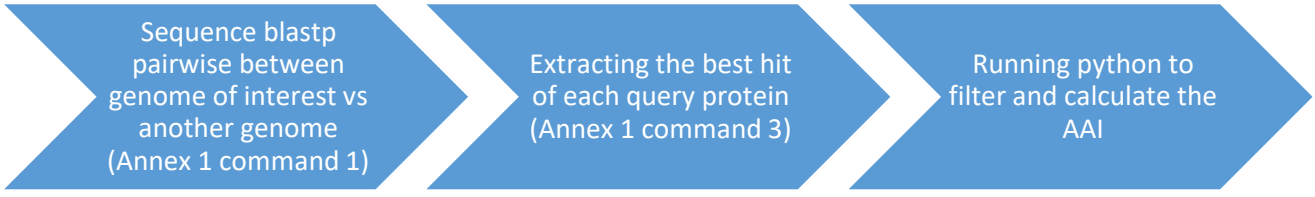


Figure 3: Steps of calculating AAI (commands in annex)

2.3.3. Calculating AF

Alignment fraction is a measure of the percentage of a DNA or protein sequence that can be aligned with another sequence. It is calculated by dividing the length of the aligned portion of the sequence by the total length of the sequence, and is typically expressed as a percentage. A higher alignment fraction indicates a higher level of sequence similarity between the two sequences, while a lower alignment fraction indicates a lower level of sequence similarity.

$$AF = \frac{\sum length\ of\ BH\ genes}{\sum length\ of\ genes\ in\ genome\ of\ interest}$$

2.3.4. Calculating the frequency of usage of codons

Using the CUSP tool available on EMBOSS[11]. It is used for understanding the patterns of codon usage in a particular gene, and for identifying potential biases in codon usage that may be relevant for studies of gene expression or protein translation. In our approach, it will be employed to compare between the five species of our study.

2.3.5. Calculating CAI

CAI (Codon Adaptation Index) is a measure of the degree to which the codon usage of a particular DNA or RNA sequence is biased towards those codons that are frequently used in a particular species. It's a value that varies between 0 and 1. A higher CAI value indicates that the codon usage of the sequence is more biased towards the optimal codons and is therefore likely to be more efficiently translated. We will also use EMBOSS's CAI[11] tool.

D. RESULTS

1. Statistics

By using data mining functions with Python, or simple NCBI searches, the following statistics were generating (table 5).

	Collinsella aerofaciens (GCA_003856815.1)	Collinsella provencensis (GCF_900199705.1)	Collinsella intenstinalis (GCF_902501455.1)	Senegalimassilia anaerobia (GCF_000236865.1)	Senegalimassilia faecalis (GCF_004135645.1)
Sequence length (bp)	2,332,121	1,737,929	1,776,933	2,398,052	2,748,043
%GC	59.7	58.2	62.4	61.8	61.2
Number of contigs	1	11	30	84	2
Protein Count	1934	1436	1522	1904	2211
Rrna	15	6	11	3	21
Trna	58	49	52	56	52
Other rna	3	3	3	3	4
Genes	2031	1518	1597	1,990	2312
Pseudogene	21	24	9	25	24
Largest protein sequence	10791	2253	4322	2011	26543
Transcriptional units (Number of operons)	5, 5, 5 (5S, 16S, 23S)	1, 3, 2 (5S, 16S, 23S)	1, 5, 5 (5S, 16S, 23S)	1, 1, 1 (5S, 16S, 23S)	7, 7, 7 (5S, 16S, 23S)
Max CDS coding protein	10791	6036	12969	14025	13509

AVG CDS coding protein	1028.238366	1035.962116	1016.683311	1038.69958	1032.846916
SUM CDS coding protein	1988613	1586058	1547392	1977684	2293953
Max genes	10791	6036	12969	14025	13509
AVG genes	997.7444609	1005.126408	976.5819775	1005.31994	1008.894488
SUM genes	2026419	1606192	1560578	2001592	2342653
Coding density	85.27057558	91.26138064	87.08218036	82.47043851	83.4758772

Table 5: Comparative statistics for studying the five species. These statistics include genes, proteins, RNAs, genomic characteristics...

These species have similar %GC content, the number of contigs depend highly on the assembly level of these genomes. What is quite different is the number of ribosomal and transfer RNAs that are found in them. The RNAs components are identical (5S, 16S, 23S) but their number differ from one to another.

2. Results on the genome scale

2.1. Dinucleotide frequency

Looking at figure 4, the observed frequency of dinucleotide occurrences is divided by the frequency in conditions where the probability is equal. Overall, the resulting pattern of comparison of the observed/expected frequency of dinucleotide occurrences is almost identical for the five species.

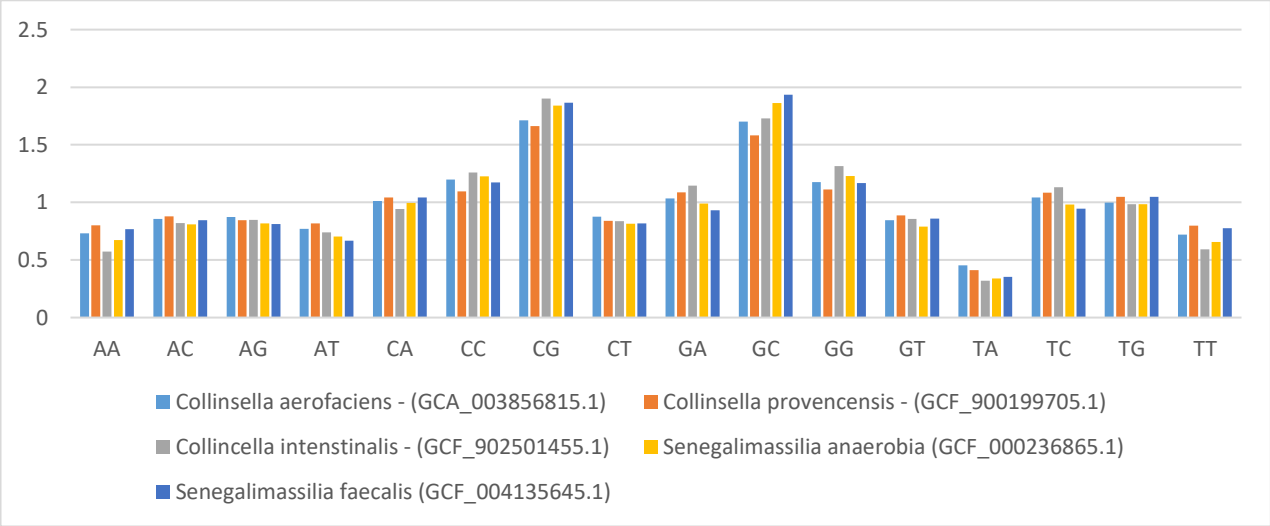


Figure 4: The distribution of the observed/expected frequency of dinucleotide occurrences in the five species

Some remarks can be made around the above “equal probability” conditions (having values > 1), as we can see that this is the case of “CG” and “GC” which reflects the high content of GC in these species.

2.2. Conserved genomic regions (D-GENIES)

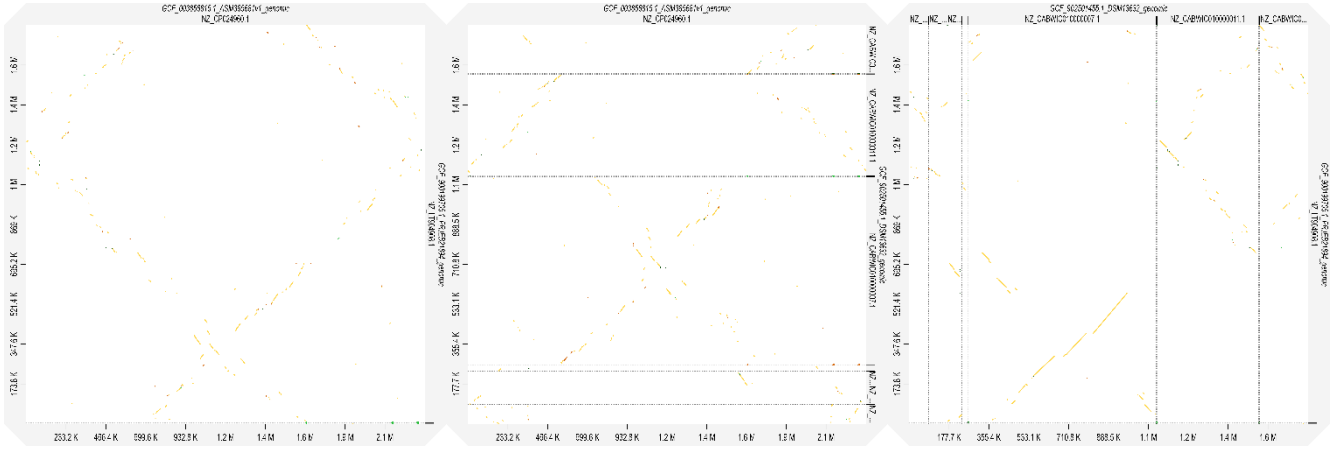


Figure 5: D-GENIES plot dot results. C. aerofaciens VS C. provencensis (LEFT), C. aerofaciens VS C. intestinalis (Middle), c. provencensis VS C. intestinalis (RIGHT)

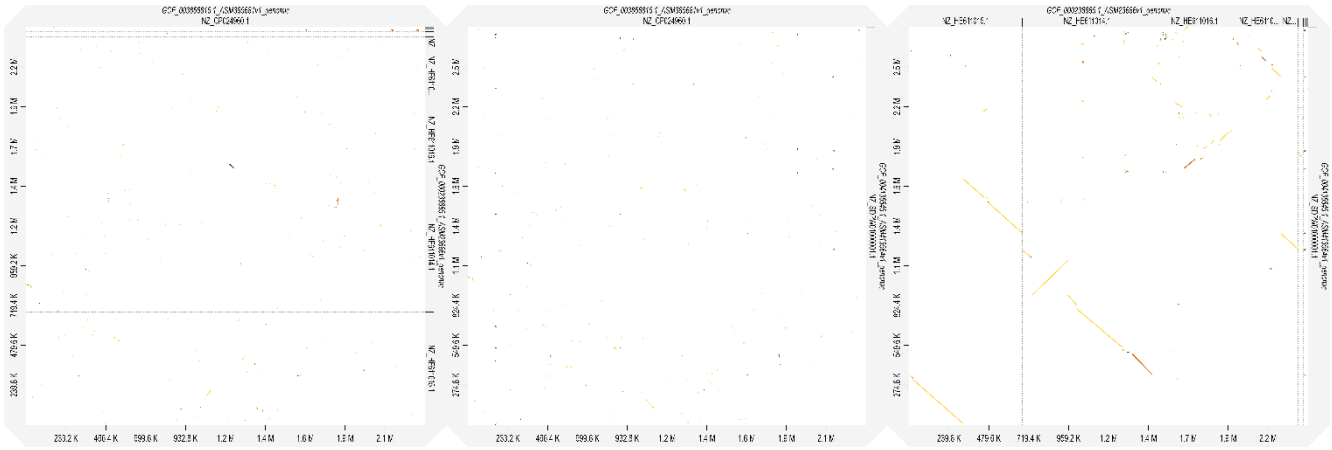


Figure 6: D-GENIES plot dot results. C. aerofaciens VS S. anaerobia (LEFT), C. aerofaciens VS S. faecalis (Middle), S. faecalis vs S. anaerobia (RIGHT)

The dot-plot nucleotide alignment of the C. aerofaciens with C. provencensis and C. intestinalis showed some regions of alignment (figure 5). However, C. aerofaciens against the Senegalimassilia species showed little to none alignments (figure 6). Alternatively, the alignment between both Senegalimassilia species shows some events of inversion and alignment that would be interesting to explore, but this subject is beyond the scope of this paper. The conserved synteny blocks between C. aerofaciens and the other two species of Collinsella are far more important than those found in the comparison with Senegalimassilia (table 7).

C. aerofaciens VS	C. provencensis	C. intestinalis	S. anaerobia	S. faecalis
Number of synteny blocks	230	284	180	67
Same direction	120	145	90	75
Opposite direction	110	139	90	92
Total size of synteny	769975	874468	327706	328716

Table 6: Table comparing the blocks of synteny

To understand the conserved sequences inside the Collinsella species, the coordinates of the alignment and role of these sequences were extracted. Quickly, the conserved sequences have in majority an enzymatic role such as the Glycosyl transferase, histidine kinase, amidohydrolase, cellulose family hydrolase, or 16s RNA-transferase and other proteins such as GnTR family transcriptional regulator.

3. Results on the proteome scale

3.1. EggNOG comparison

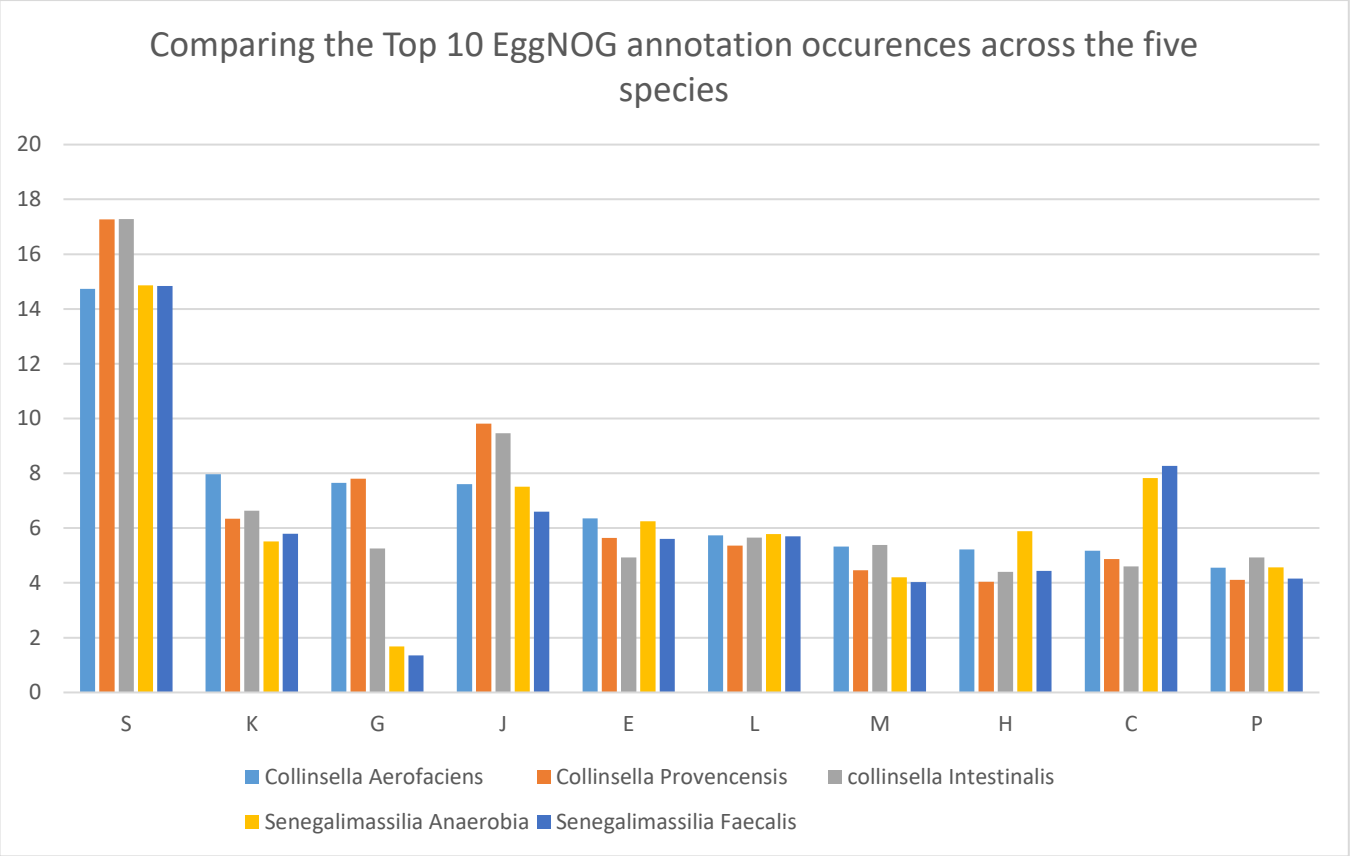


Figure 7: The percentage of frequency of the most 10 occurred OG annotations of all the species

By comparing the bars of figure 7, the frequency of the top ten OG annotations inferred by eggnog-mapper are highly similar in the percentage
Some observations can be made, higher C for Senegalimassilia implied in "energy production and conversion" and higher G for the Collinsella genus implied in "Carbohydrate transport and metabolism". Nevertheless, these two annotations are related to the metabolism activity of this family of bacteria. These results align well with what is expected from these species.
Other OG categories were redacted because they occurred only once or twice.

3.2. COG Vs EggNOG-mapper annotation

In the COG database, the query of C. aerofaciens yield the annotation of 1101 proteins out of the 1934 proteins produced by this genome (annotation rate of 56.92%) whereas EggNOG-mapper has annotated 1778 proteins in the same Proteome (annotation rate of 91.93%). The COG database annotation is mainly manually curated, hence it's expected to have better accuracy. For example, an abundant number of proteins were classified as S "Function Unknown" (figure 8), which isn't very intuitive or meaningful in this case.

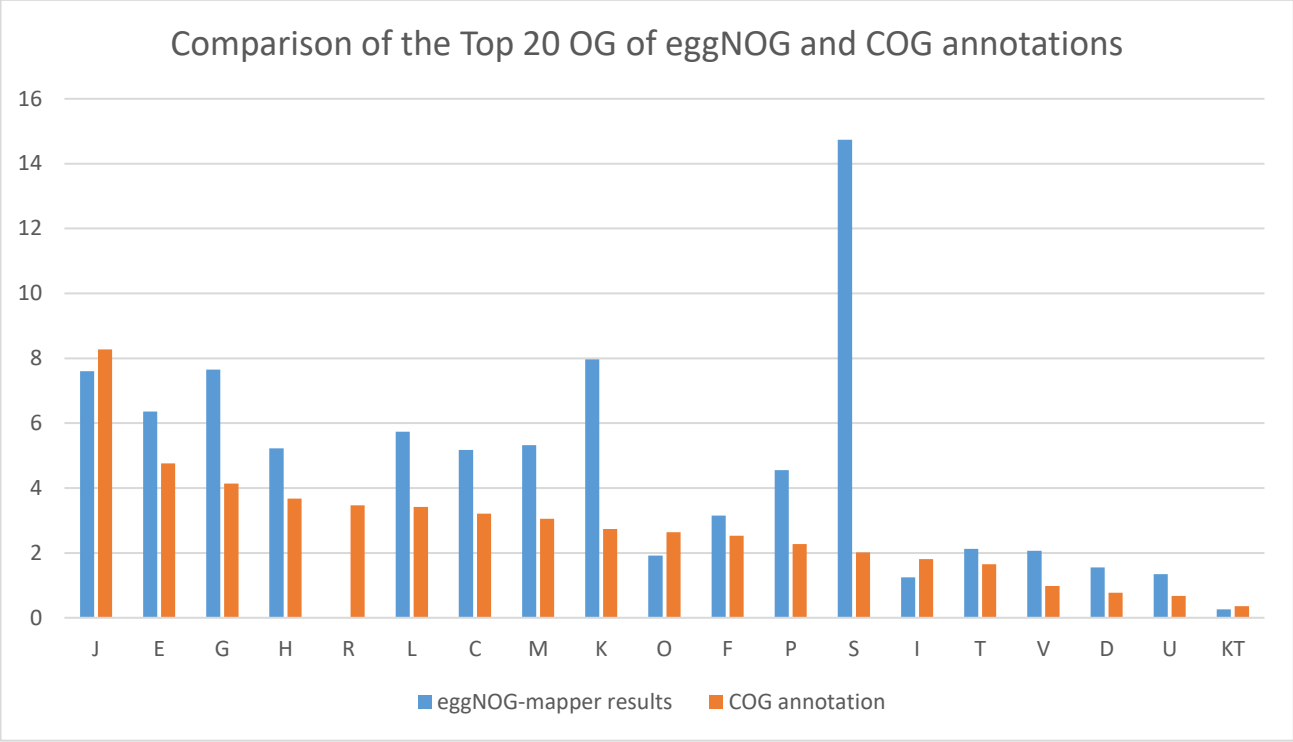


Figure 8: The percentage of frequency of the most 20 occurred eggNOG and COG annotations for C. aerofaciens

3.3. Finding Orthologous clusters with OrthoVenn 2

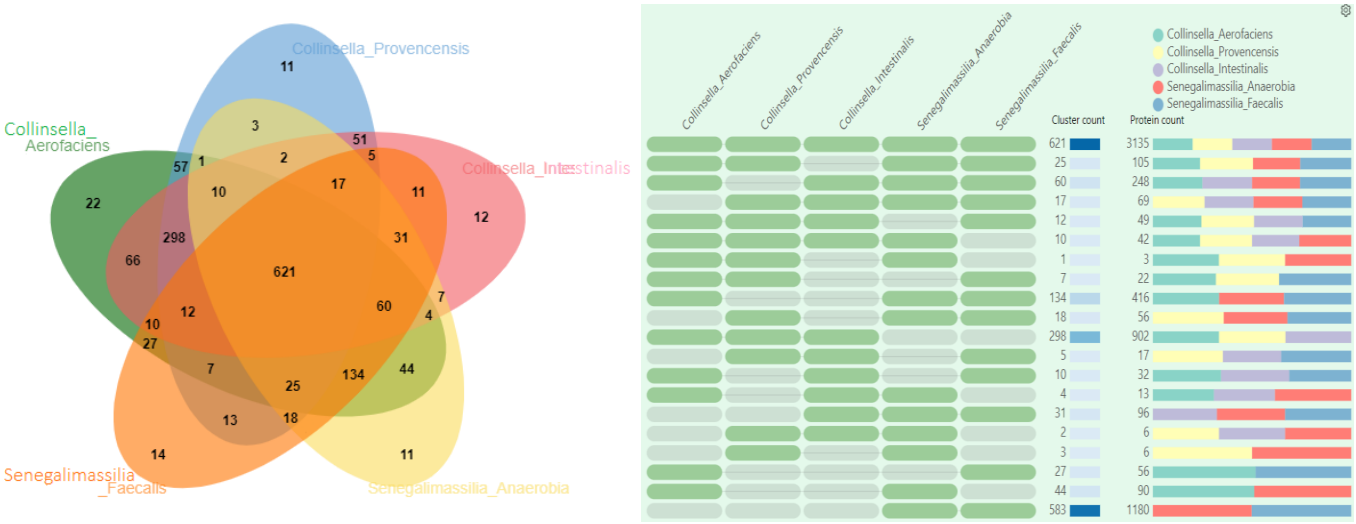


Figure 9: Venn Diagram showing the intersection of the cluster of genes of the five species (Left). Repartition of the clusters and the protein counts for the combination of species and their intersection (Right).

The results of OrthoVenn (figure 9) show that 3135 proteins are shared between the five genomes (Core genome) making up to 621 clusters. This partition is second highest for the clusters shared by Senegalimassilia species (1180 proteins – 583 clusters), then Collinsella species (902 proteins – 298 clusters). This observation confirms the idea that the closely related species in the phylogeny tree of species should share together a set of proteins having orthologous relations.

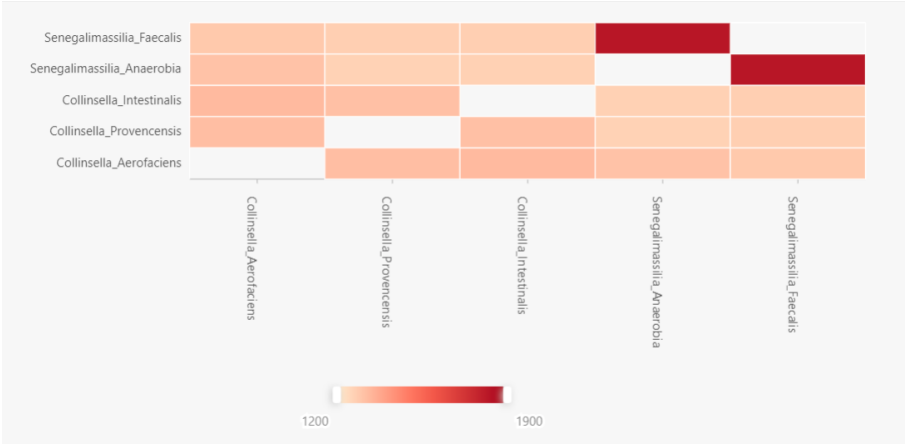


Figure 10: Pariwise similarity heatmap generated by OrthoVenn

The heatmap in figure 10 shows a good segregation of the pairwise relation between Senegalimassilia species, but it doesn't show the same intensity of similarity for the Collinsella species.
For understanding the characteristics and specifications of the C. aerofaciens specie, the extraction of the GO annotation of its singleton proteins has been done (table 8). It revealed that the proteins found to be specific for this organism are involved in the processes of polysaccharide's metabolism.

GO:0005975	carbohydrate metabolic process
GO:0005976	polysaccharide metabolic process
GO:0006139	nucleobase-containing compound metabolic process
GO:0006725	cellular aromatic compound metabolic process
GO:0006793	phosphorus metabolic process
GO:0006807	nitrogen compound metabolic process

Table 8:GO annotations of C. aerofaciens singleton proteins

3.4. Identifying Orthologous genes with phylogenetic inference (OrthoFinder)

	collinsella_aerofaciens	collinsella_intestinalis	collinsella_provencensis	senegalimassilia_anaerobia	senegalimassilia_faecalis
collinsella_aerofaciens	1384	1067	1018	993	985
collinsella_intestinalis	1067	1189	1002	830	852
collinsella_provencensis	1018	1002	1127	777	800
senegalimassilia_anaerobia	993	830	777	1536	1445
senegalimassilia_faecalis	985	852	800	1445	1539

Figure 11: Comparative statistics of the overlap of the genes of the species in the Orthogroups

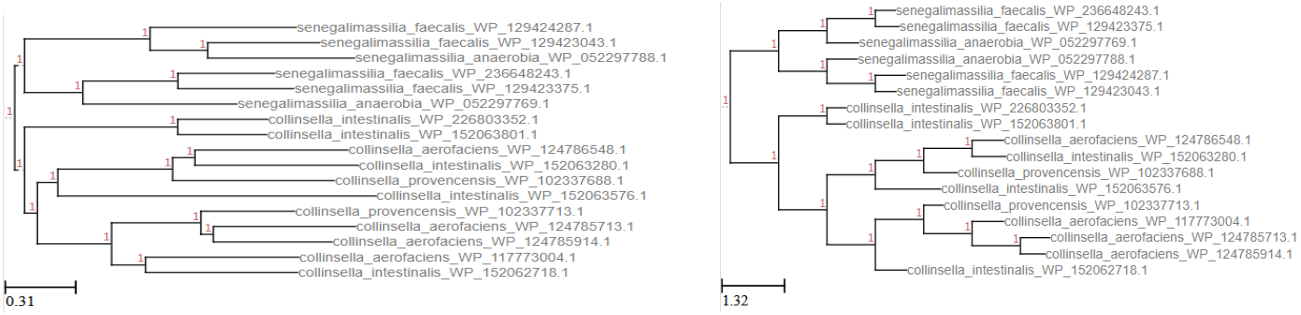


Figure 12: Example of a tree of orthologous group (number 5 in this case). Unresolved tree (Left) Resolved tree (Right).

OrthoFinder yielded more than 1015 inferred gene trees for each the OG clusters built, hence in our approach we will only focus one some of them and on the overall comparative OrthoFinder stats. It's shown that the Collinsella species share among them a large quantity of orthologous proteins that is higher than the number that they share with the Senegalimassilia genus (figure 11). The same observation is made for Senegalimassilia species as they share among each other biggest number of overlapping proteins. Again, this consolidates the consensus that the

closer the species are, the more similarity they will have in their proteins due to the resemblance in functions and pathways.
From Figure 12 the resulting gene tree of one of the orthologous groups (group 5 of our data) show proteins that all belong to the same orthologous group based on the Orthofinder classification. The two genus of species are nicely separated. Also by comparing both of the phylogenetic trees, it's possible to observe the differences between the unresolved gene tree and the resolved one.

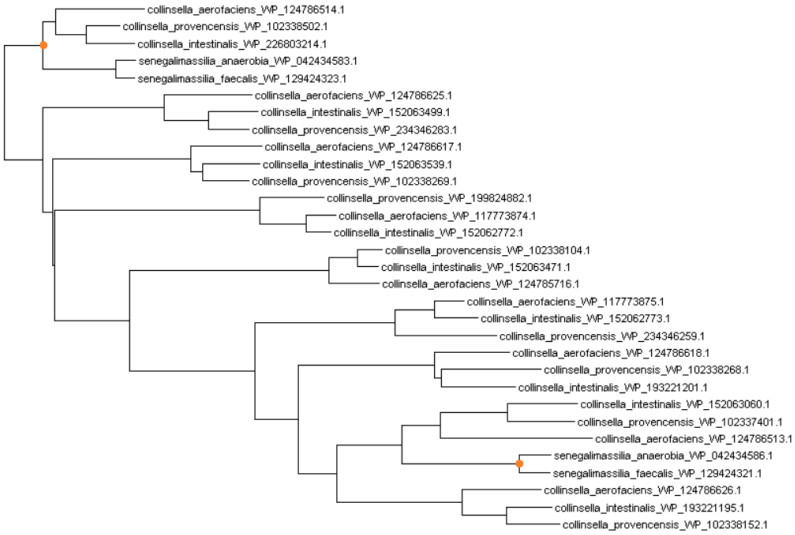


Figure 13: Phylogram of Orthologous group 0

Figure 13 shows another example result is cluster 0 which groups proteins with "ABC transporter ATP-binding protein" function. This orthologous group present a collection of protein across the five species that have the same role whereas the closely related species are classified together, but this is not the case all the time, hence the phylogram reveals a certain possibility of duplication events across species. This is the interest of the usage of the phylogenetic tree in indicating possible events that occurred throughout evolution between the species.

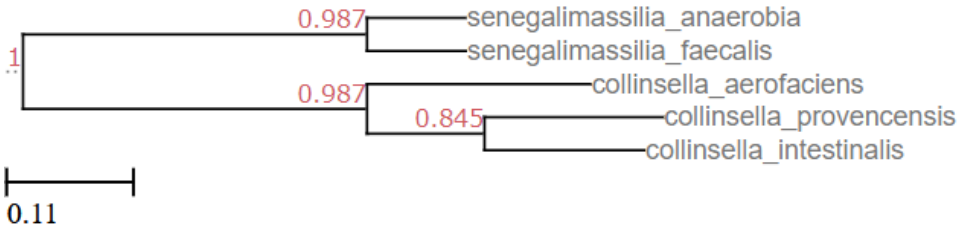


Figure 14: Rooted species tree inferred by Orthofinder

The inferred species tree based on the orthology relationships between the proteins is shown in figure 14. Based on our prior knowledge of the species selected, this classification is right as it groups the species belonging to the same genus together. Furthermore, C. provencensis and C. intestinalis are more closely related to each other than they are to C. aerofaciens.

4. Results on the scale of the coding sequences

4.1. Comparing the frequency of codon usage (Cusp)

The frequency of usage of each codon is very similar between the five species as seen on figure 15. The patterns are easily identifiable when they are compared together. Nevertheless, some differences arise. For example, higher usage of TTT for the C. aerofaciens compared to the other species, higher usage of GTG for the Senegalimassilia species compared to Collinsella.

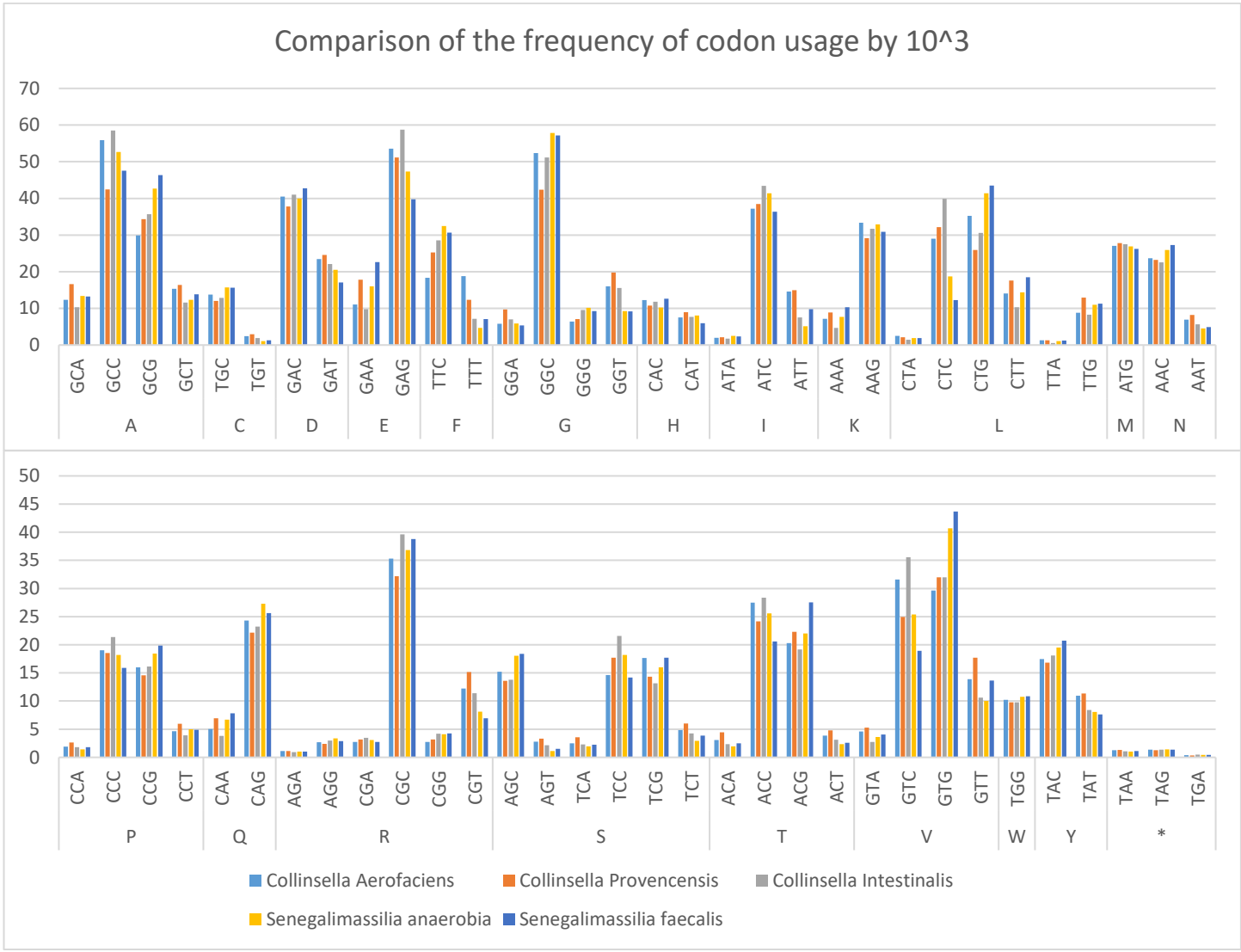


Figure 15: Charts comparing the frequency of the codon usage by 10^3 with the corresponding amino acid beneath

This result is also confirmed by the codon usage per amino acid shown in Annex 2.

4.2. Alignment of the coding sequences (ANI, AFI, AI)

Looking at figure 16, the results of comparison of AF between C. aerofaciens and each species show that AF is the highest in C. intestinalis then C. provencensis and finally the Senegalimassilia species. It's a similar case for ANI and AAI, the same observation is made. This shows that there is a higher similarity of both protein and DNA sequences of the C. aerofaciens with the species that belong to the same genus. However, the results with Senegalimassilia aren't "extremely" low, which means that there exists a certain similarity between them and C. aerofaciens. This is normal for the fact that these species belong to the same family; a certain selection is done to preserve for example a metabolic pathway or particular functionality that is vital for these species.

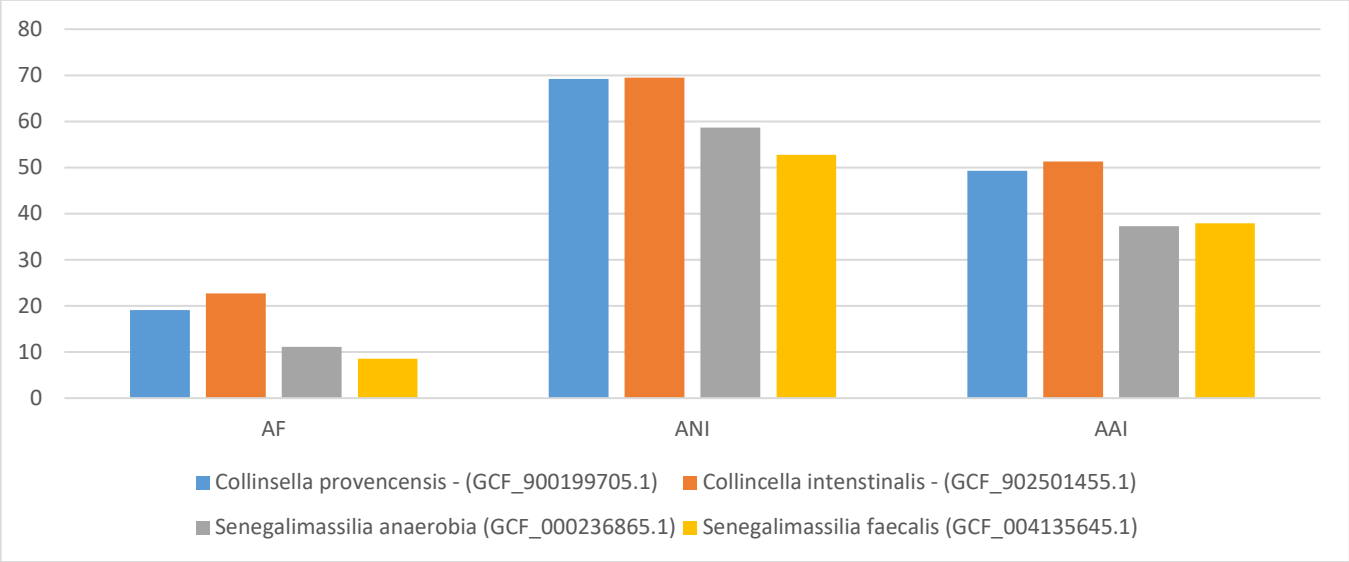


Figure 16: Comparison of the AF, ANI and AAI between the five species

4.3. Studying the adaptation of the codons (CAI)

The CAI scores of all CDS of the genome were sorted from the biggest value to the smallest and plotted as seen in figure 17. For each specie, the score starts with the highly adapted genes of the genome. Around the 900th CDS, we can start to see some differences between the scores across the species. However, it can be seen that the CAI of the Senegalimassilia CDS are higher than those of Collinsella. For example, some of the values of the CAI of Collinsella can reach 0.28.

That means that the proteins of these genomes don't necessarily present the same transcription activity, as we have seen overall the genes of Senegalimassilia should present, on average, a higher activity.

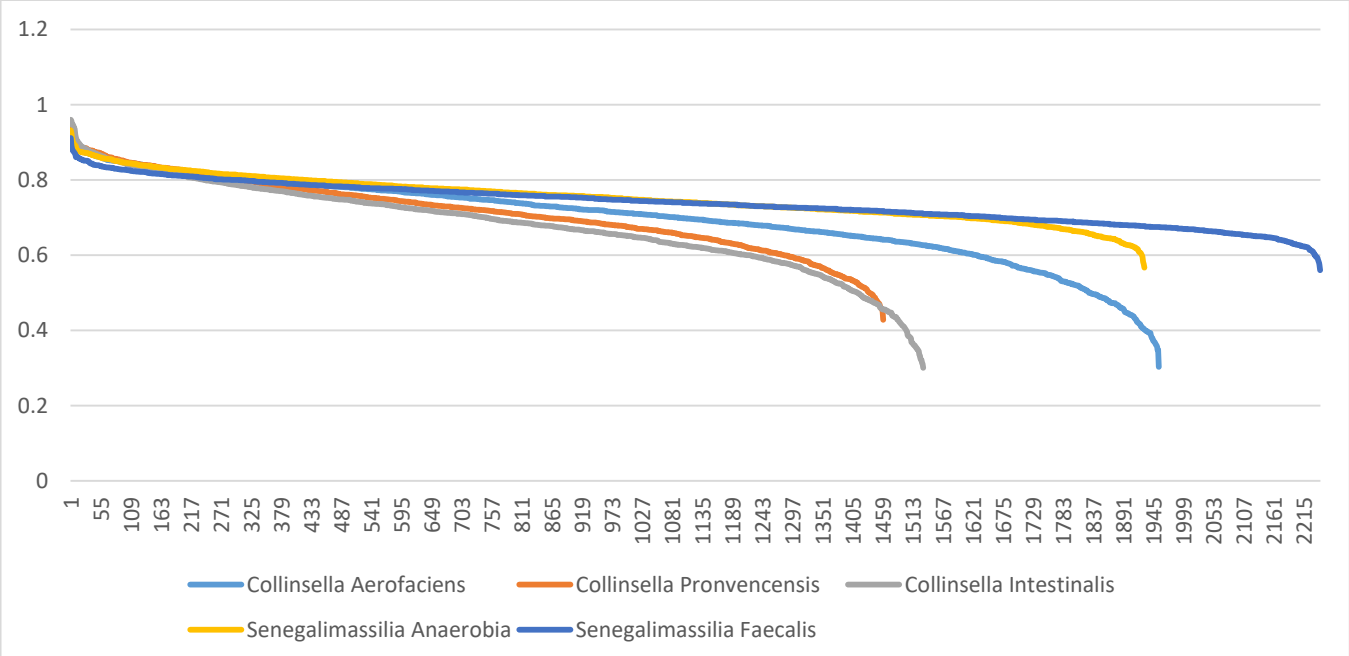


Figure 17:CAI of the CDS of all five spcies sorted in ascending order

E. DISCUSSION AND CONCLUSION

After analyzing all five species that belong to Coriobacteriaceae's family with three of them that belong to the genus Collinsella and two that belong to Senegalimassilia, on different levels, show that these species share among them several common points and pathway. This result was expected since all these species belong to the intestine's microbe gut and have a similar general role of sugar degradation.

The genome comparison investigation offered some intriguing into the evolutionary links and distinctive traits of Collinsella aerofaciens in compared with other Collinsella species. The identification of special genetic markers or characteristics that are exclusive to C. aerofaciens, such as metabolic pathways that enable it to flourish in various conditions, could be one of the study's most important findings. The investigation may have also identified shared genetic traits or ancestry among the various Collinsella species, adding to the evidence of those links.
The alignment sequence observed on D-GENIES indicates that a collection of the genes coding for certain enzymes are preserved in the Collinsella's genus. Hence, a collection of functional enzymes and pathways are preserved between the closely-related species of the family of Coriobacteriaceae. As we have observed using the comparative approach, certain signatures differentiate between the c.aerofaciens and other species that belong to the same genus. More investigation into these bacteria's genetic traits may help us comprehend their functions in various ecological systems. Hence, it might be also useful to choose other species that may belong to the family of Coriobacteriaceae but that are physiologically and ecologically different than C. aerofaciens to eliminate any possible bias do to the fact that the Senegalimassilia species that were chosen aren't very different from the Collinsella. For example, the Olegusella Massilienis belongs to Coriobacteriaceae but shows more differences toward C. aerofaciens (one of them having %GC <50 (Annex 3)).

F. REFERENCES

[1] J. C. Setubal, N. F. Almeida, and A. R. Wattam, "Comparative Genomics for Prokaryotes," *Methods Mol Biol*, vol. 1704, pp. 55–78, 2018, doi: 10.1007/978-1-4939-7463-4_3.

[2] N. Dione *et al.*, "'Collinsella provencensis' sp. nov., 'Parabacteroides bouchesdurhonensis' sp. nov. and 'Sutterella seckii,' sp. nov., three new bacterial species identified from human gut microbiota," *New Microbes New Infect*, vol. 23, pp. 44–47, May 2018, doi: 10.1016/j.nmni.2018.02.003.

[3] A. Kageyama and Y. Benno, "Emendation of genus Collinsella and proposal of Collinsella stercoris sp. nov. and Collinsella intestinalis sp. nov," *Int J Syst Evol Microbiol*, vol. 50 Pt 5, pp. 1767–1774, Sep. 2000, doi: 10.1099/00207713-50-5-1767.

[4] J.-C. Lagier, K. Elkarkouri, R. Rivet, C. Couderc, D. Raoult, and P.-E. Fournier, "Non contiguous-finished genome sequence and description of Senegalemassilia anaerobia gen. nov., sp. nov," *Stand Genomic Sci*, vol. 7, no. 3, pp. 343–356, 2013, doi: 10.4056/sigs.3246665.

[5] K.-I. Han *et al.*, "Senegalimassilia faecalis sp. nov., an anaerobic actinobacterium isolated from human faeces, and emended description of the genus Senegalimassilia," *Int J Syst Evol Microbiol*, vol. 70, no. 3, pp. 1684–1690, Mar. 2020, doi: 10.1099/ijsem.0.003958.

[6] R. L. Tatusov, M. Y. Galperin, D. A. Natale, and E. V. Koonin, "The COG database: a tool for genome-scale analysis of protein functions and evolution," *Nucleic Acids Research*, vol. 28, no. 1, pp. 33–36, Jan. 2000, doi: 10.1093/nar/28.1.33.

[7] S. Astbury, E. Atallah, A. Vijay, G. P. Aithal, J. I. Grove, and A. M. Valdes, "Lower gut microbiome diversity and higher abundance of proinflammatory genus Collinsella are associated with biopsy-proven nonalcoholic steatohepatitis," *Gut Microbes*, vol. 11, no. 3, pp. 569–580, May 2020, doi: 10.1080/19490976.2019.1681861.

[8] A. H. Eggerth, "The Gram-positive Non-spore-bearing Anaerobic Bacilli of Human Feces," *J Bacteriol*, vol. 30, no. 3, pp. 277–299, Sep. 1935, doi: 10.1128/jb.30.3.277-299.1935.

[9] A. Kageyama, Y. Benno, and T. Nakase, "Phylogenetic and phenotypic evidence for the transfer of Eubacterium aerofaciens to the genus Collinsella as Collinsella aerofaciens gen. nov., comb. nov," *Int J Syst Bacteriol*, vol. 49 Pt 2, pp. 557–565, Apr. 1999, doi: 10.1099/00207713-49-2-557.

[10] N. A. O'Leary *et al.*, "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation," *Nucleic Acids Res*, vol. 44, no. D1, pp. D733-745, Jan. 2016, doi: 10.1093/nar/gkv1189.

[11] P. Rice, I. Longden, and A. Bleasby, "EMBOSS: The European Molecular Biology Open Software Suite."

[12] F. Cabanettes and C. Klopp, "D-GENIES: dot plot large genomes in an interactive, efficient and simple way," *PeerJ*, vol. 6, p. e4958, 2018, doi: 10.7717/peerj.4958.

[13] D. M. Emms and S. Kelly, "OrthoFinder: phylogenetic orthology inference for comparative genomics," *Genome Biology*, vol. 20, no. 1, p. 238, Nov. 2019, doi: 10.1186/s13059-019-1832-y.

[14] C. P. Cantalapiedra, A. Hernández-Plaza, I. Letunic, P. Bork, and J. Huerta-Cepas, "eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale," *Mol Biol Evol*, vol. 38, no. 12, pp. 5825–5829, Dec. 2021, doi: 10.1093/molbev/msab293.

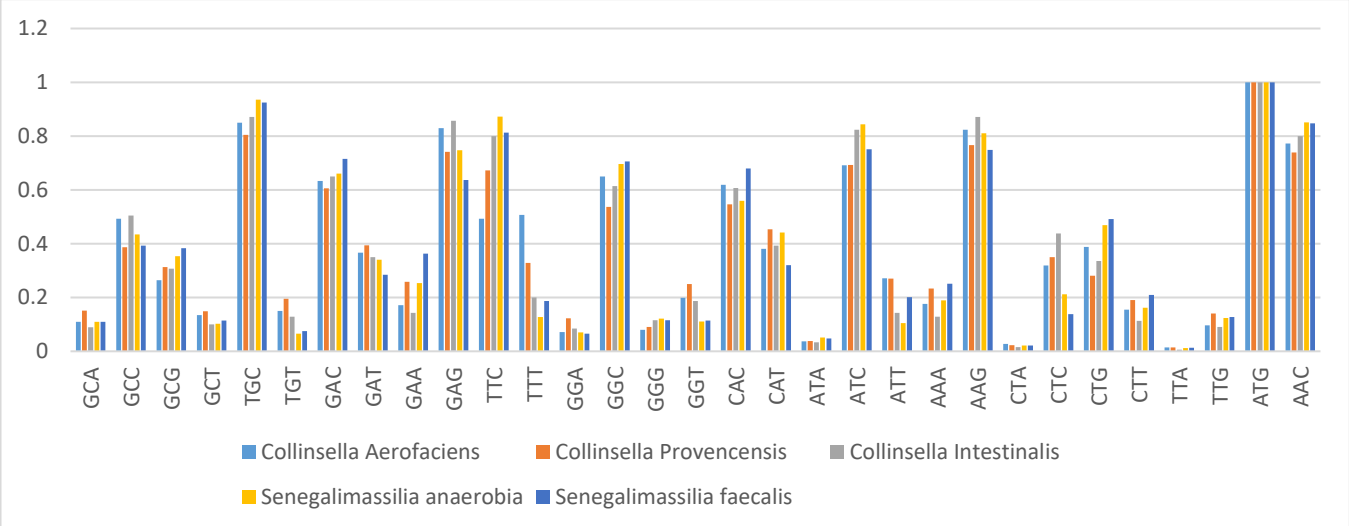
[15] L. Xu *et al.*, "OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species," *Nucleic Acids Res*, vol. 47, no. W1, pp. W52–W58, Jul. 2019, doi: 10.1093/nar/gkz333.
[16] D. M. Emms and S. Kelly, "OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy," *Genome Biology*, vol. 16, no. 1, p. 157, Aug. 2015, doi: 10.1186/s13059-015-0721-2.

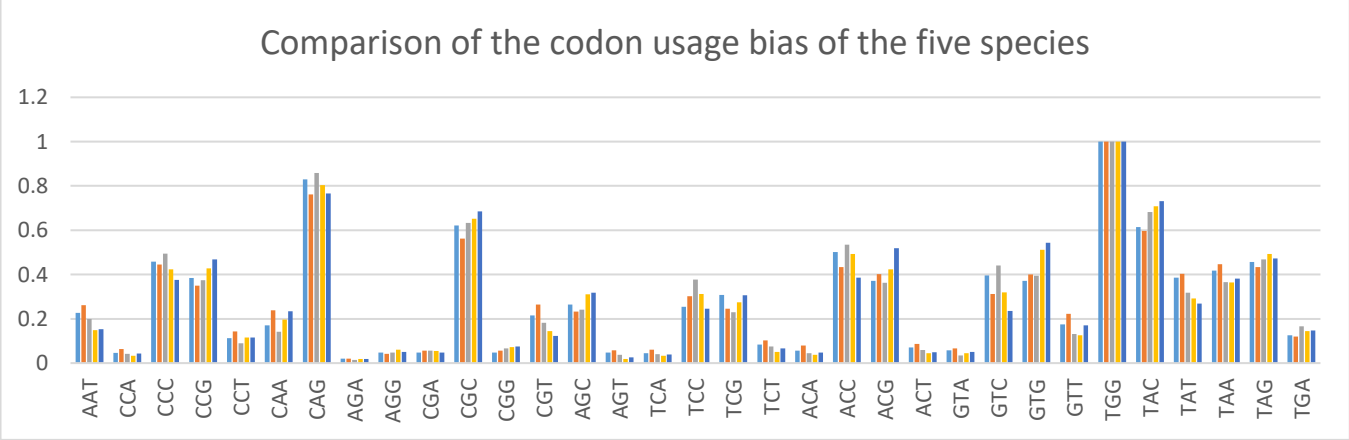
G. Annex

Additional data and materials are available at: https://github.com/GeorgeAlehandro/GECOProject

Command number	Bash commands
1	blastp -query {query.faa} –subject {subject.faa} -out {file} -outfmt "6 qseqid sseqid qlen evalue score pident qcovs length"
2	blastn -query {query_cds_from_genomic.fna} -subject {subject_cds_from_genomic.fna} -out {file}-outfmt "6 qseqid sseqid qlen slen evalue score pident qcovs length"
3	for next in $(cut -f1 {file} \| sort -u); do grep -w -m 1 "$next" {file}; done > {BH_file}
4	for f in collinsella_aerofaciens.faa collinsella_provencensis.faa collinsella_intestinalis.faa senegalimassilia_anaerobia.faa senegalimassilia_faecalis.faa ; do python ~/OrthoFinder/tools/primary_transcript.py $f ; done
5	~/OrthoFinder/orthofinder -f primary_transcripts

Annex 1: Table of Bash commands





Annex 2: Comparison of the codon usage bias per amino acid across all species

	Olegusella massiliensis (GCF_900078545.1)	Parvibacter caecicola (GCF_005046025.1)
Sequence length (bp)	1,806,744	2,412,878
%GC	49.2	62.4
Protein Count	1531	1839
Rrna	6	3
Trna	45	46
Other rna	3	3
Genes	1598	1922
Pseudogene	13	31
Largest protein sequence	2011	26543
Unites transcriptionelles (Nombre des operons)	2, 2, 2 (5S, 16S, 23S)= 2	1, 1, 1 (5S, 16S, 23S)
Max CDS coding protein	6762	79632
AVG CDS coding protein	1016.978412	1134.171289
SUM CDS coding protein	1460381	2085741
Max genes	6762	79632
AVG genes	977.6561265	1099.583247
SUM genes	1484082	2113399
Coding density	80.82943682	86.44204141

Annex 3: Table of statistics for two more species that belong to the same family

Name of the file	Brief description
scriptsReadingBestHit.py	Filtering of the BH for calculating AAI, AF, ANI
scriptsReadingDgeniesAlignmentRegions.py	Retrieving coordinates of paf files and fetchin the corresponding features annotated in these regions
stats_scripts.py	Generating statistics out of feature tables
GECO_scripts.ipynb	Groups scriptsReadingBestHit.py and stats_scripts.py in one jupyter notebook

Annex 4: Table of the python scripts files (please note that depending on the OS of the machine or the directory, some modifications could be needed)

Name of the file	Content
cai_comparison_five_species.xslx	Contains the CAI score of all proteins of the species sorted from the biggest to the smallest value
comparison_cog_eggnog_c_aerofaciens.xlsx	Compares between the COG annotation and the eggnog-mapper results categories
comparison_eggNOG_5_species.xlsx	Compares the categories and counts of eggnog annotation of the five species
Comparative_statistics.xlsx (various sheets inside)	Various comparative statistics
	Dinucleotide frequencies
	AF ANI AAI
	Compseq results extraction

	Cumulative frequency Cusp
	Cusp statistics

Annex 5: Additional description of excel files of the repository