UPPSALA
UNIVERSITET

UPTEC X 21018
Examensarbete 30 hp
Juni 2021

Study of the Genetic Dynamics in Pan-genomes for Six Bacterial Species

Jennifer Johansson

UPPSALA UNIVERSITET

Teknisk- naturvetenskaplig fakultet
UTH-enheten

Besöksadress:
Ångströmlaboratoriet
Lägerhyddsvägen 1
Hus 4, Plan 0

Postadress:
Box 536
751 21 Uppsala

Telefon:
018 – 471 30 03

Telefax:
018 – 471 30 00

Hemsida:
http://www.teknat.uu.se/student

Abstract

Study of the Genetic Dynamics in Pan-genomes for Six Bacterial Species

Jennifer Johansson

Foodborne diseases are a growing health problem today and can be caused by eating food contaminated with bacteria. To monitor known foodborne diseases, institutions keep track of bacteria in surveillance projects. Whole genome sequencing is becoming the new standard method for comparing isolates, which generates large amounts of data. Today, the standard analyses are focused on conserved regions in genomes. The dynamics in less conserved regions can be studied by creating pan-genomes. A pan-genome consists of conserved genes, called core genes, and genes of varied conservation grade, called accessory genes. This thesis aimed to analyse pan-genomes of large datasets from six bacterial species coming from surveillance projects: *Campylobacter coli*, *Campylobacter jejuni*, *Escherichia coli*, *Listeria monocytogenes*, *Salmonella enterica*, and *Streptococcus pneumoniae*. The purpose was to investigate the species dynamics in the genomes and to look at properties of the genomes not included in the standard analyses that are used in surveillance projects today.

Bacterial Pan Genome Analysis tool was used for the pan-genome analysis of the six species and datasets of 1,000-2,000 genomes per species were analysed. All species were estimated to have open pan-genomes, meaning the pan-genomes are increasing in size as more genomes are added. *Escherichia coli* and *Salmonella enterica* had more dynamic and open genomes compared to the other species. They had the highest number of accessory genes relative to their genome sizes and had the largest accessory segments between core genes. The synteny of the core genes showed high conservation for a part of the core genes in all species. Some core genes always sat directly after each other in the analysed genomes, never having accessory genes between them. Other core genes always had accessory genes between them, indicating very open regions in the genomes. The core genes were evenly distributed through the reference genomes with some regions showing increased gene density for all species. Some regions had a higher gene density for core genes often followed by core genes, and others for core genes often followed by accessory genes. However, the placement of genes needs to be investigated further with more reference genomes to be able to draw confident conclusions.

Handledare: Bo Segerman
Ämnesgranskare: Dan I. Andersson
Examinator: Pascal Milesi
ISSN: 1401-2138, UPTEC X 21018

Sammanfattning

Varje år insjuknar ungefär 10 % av världens befolkning i livsmedelsburna sjukdomar. Detta är ett växande problem och det finns idag mer än 200 kända sjukdomar som orsakas av att människor äter kontaminerad mat. Varje år sker fler än 420 000 dödsfall till följd av livsmedelsburna sjukdomar och av dessa är 125 000 barn yngre än fem år. Sjukdomarna kan bland annat orsaka diarré och cancer och orsakas av att maten vi äter är kontaminerad av olika kemikalier, bakterier, virus och parasiter.

För att övervaka och studera livsmedelsburna sjukdomar utför olika institutioner över hela världen övervakningsprojekt där bland annat bakterier studeras. Statens Veterinärmedicinska Anstalt (SVA) i Uppsala är ett av de ställen som jobbar med olika övervakningsprojekt. Syftet med övervakningsprojekten är bland annat att detektera utbrott av livsmedelsburna sjukdomar, studera varför utbrott sker och att i möjligaste mån minska effekterna av pågående utbrott. Detta kan till exempel göras genom att identifiera smittkällor.

I det här projektet analyserades de sex vanligaste bakterierna som studeras i övervakningsprojekt: *Campylobacter coli*, *Campylobacter jejuni*, *Escherichia coli*, *Listeria monocytogenes*, *Salmonella enterica* och *Streptococcus pneumoniae*. Syftet var att studera de olika bakteriernas gener och undersöka dynamiken i deras genom. Ett annat syfte var att undersöka hur många bakterier från varje bakterieart som kunde analyseras. En hypotes som undersöktes i projektet var ifall gener som bara finns i vissa bakterier hos en population sätter sig på specifika platser i genomen eller om de slumpmässigt placeras i genomen.

Det första steget i projektet var att skapa pan-genom för de olika bakteriearterna, vilket visar hur många gener varje art totalt sett innehåller utifrån de bakterier som analyseras. Detta gjordes med hjälp av det bioinformatiska verktyget Bacterial Pan Genome Analysis tool där bakteriernas gener delades upp i olika kategorier som alla är del av pan-genomet:

- Coregener: gener som fanns hos alla bakterier
- Accessorygener: gener som inte fanns hos alla bakterier men hos minst två
- Unika gener: gener som endast fanns hos en bakterie

Utöver detta gjorde programmet en uppskattning för hur öppna de olika arterna är för att ta upp nya gener från omgivningen. Efter detta gjordes ytterligare analyser som undersökte hur konserverade coregenernas placering i genomen är och hur ofta olika coregener tillåter att accessorygener sitter mellan specifika coregener. För de platser i genomen där coregenerna tillät att accessorygener satt mellan dem analyserades hur långa accessory-gensegment som satt mellan coregenerna. Till sist användes ett referensgenom för varje art att jämföra

coregenerna mot och därmed få en uppskattning kring hur generna var fördelade över referensgenomen.

Pan-genom kunde göras för 1 000 bakterier var för *Escherichia coli* och *Salmonella enterica*. Hos de övriga arterna, *Campylobacter coli*, *Campylobacter jejuni*, *Listeria monocytogenes* och *Streptococcus pneumoniae,* kunde istället pan-genom göras för 2 000 bakterier per art. Resultatet av detta projekt visade att två arter, *Escherichia coli* och *Salmonella enterica*, hade mer dynamiska genom och är mer benägna att ta upp nya gener från omgivningen. Detta eftersom de hade ett större antal accessorygener och unika gener i förhållande till deras genomstorlek jämfört med de andra arterna. De hade även större genomsegment med accessorygener placerade mellan coregener än de andra arterna och uppskattades att ha de mest öppna genomen av Bacterial Pan Genome Analysis tool. Coregenernas placering i förhållande till varandra visade sig vara mycket konserverade för en del coregener. I en del fall följde två coregener alltid varandra i alla bakteriers genom som studerades under projektet. En del coregener satt dessutom alltid direkt efter varandra i genomen och lät inga andra gener sitta mellan dem. Detta tyder på att en del coregener är placerade i gensegment där alla gener placerade i just det segmentet uttrycks tillsammans. Andra coregener hade alltid accessorygener placerade mellan dem, vilket visade på mer dynamiska och öppna delar av genomet. I dessa delar av genomen har bakterierna lättare för att ta upp nya gener från omgivningen. När coregenerna jämfördes mot referensgenom visade det sig att de flesta generna var placerade jämnt över hela referensgenomen. Men det fanns också vissa delar i genomen där fler coregener var placerade. Det fanns också delar i genomen med antydningar att fler accessorygener kan vara placerade där. Däremot skulle ytterligare analyser med ett flertal referensgenom per art behöva göras för att kunna dra säkra slutsatser gällande genernas placering.

Sammanfattningsvis visade resultaten i det här projektet att en del bakteriearter har mer dynamiska och öppna genom än andra. Olika delar av genomen inom en art kan också vara mer eller mindre dynamiska då en del gensegment är mer benägna att släppa in nya gener. Detta leder till att nya gener inte endast fördelar sig slumpmässigt över genomen, utan det går att se vilka områden som tenderar att tillåta nya gener i högre utsträckning.

Table of Contents

Abbreviations

BPGA	Bacterial Pan Genome Analysis tool
C. coli	*Campylobacter coli*
C. fetus	*Campylobacter fetus*
C. jejuni	*Campylobacter jejuni*
E. coli	*Escherichia coli*
L. monocytogenes	*Listeria monocytogenes*
MLST	Multi-Locus Sequence Typing
NCBI	National Center for Biotechnology Information
OriC	Origin of Replication
RAM	Random Access Memory
S. enterica	*Salmonella enterica*
S. pneumoniae	*Streptococcus pneumoniae*
SNPs	Single-Nucleotide Polymorphisms
SVA	Statens Veterinärmedicinska Anstalt
Ter	Terminus of Replication
WHO	World Health Organization

1 Introduction

Almost 10 % of the world's population become ill every year from consuming contaminated food, and the World Health Organization (WHO) has stated that foodborne diseases are a growing health problem (WHO 2021). Today, over 200 known foodborne diseases have an impact on disease and mortality globally (WHO 2021). Furthermore, more than 420,000 deaths are caused by foodborne diseases each year, and out of these, 125,000 are children under five years old (WHO 2021). The diseases can be all from diarrhoea to cancer and are caused by food contaminated by chemicals, bacteria, viruses, or parasites (WHO 2021). Surveillance projects are performed to detect outbreaks of foodborne diseases as well as raise the knowledge of why outbreaks happen, predict the impact of an outbreak, and limit the impact of an ongoing outbreak (Deng *et al.* 2016).

Whole-genome sequencing is widely used today and is a standard method for analysis in many microbiological research studies. The use of it is also increasing for outbreak investigations and for monitoring pathogenic bacteria (Deng *et al.* 2016). However, today the standard analysis for bacterial outbreak investigations and surveillance projects is mainly focused on Single-Nucleotide Polymorphisms (SNPs) and extended Multi-Locus Sequence Typing (MLST) approaches utilizing the core genome. Often it can be complemented with a targeted search for known resistance genes and virulence factors. Among the methods used today, there is very limited availability of standardized methods to look at the dynamics outside of the core genome.

1.1 Project Aim

This thesis was performed at Statens Veterinärmedicinska Anstalt (SVA) in Uppsala. The purpose was to study the pan-genome structure of bacteria that are sequenced in surveillance projects by dividing their genomes up into core-, and accessory genes. Furthermore, the purpose was to better understand the dynamics of the gene contents in bacterial genomes coming from surveillance projects. The purpose was also to analyse parts and properties of the genomes that are not included in the standard core genome analysis. To do this, I chose to work with the six bacterial species which were most frequently occurring in a database composed of genomes from surveillance projects, and analyse the dynamics in the chosen species core-, and accessory genomes. One aim was to explore how large datasets could be used for the different species to successfully perform the pan-genome analysis. A hypothesis that was investigated through this project was if there are hotspots in the genomes where accessory genes are placed, or if they are placed in the genome by random. Six bacteria species were analysed in this thesis: *Campylobacter coli (C. coli)*, *Campylobacter jejuni (C. jejuni)*, *Escherichia coli (E. coli)*, *Listeria monocytogenes (L. monocytogenes)*, *Salmonella enterica (S. enterica)*, *Streptococcus pneumoniae (S. pneumoniae)*. The chosen species are the major foodborne pathogens monitored in surveillance projects and outbreak investigations today (Deng *et al.* 2016).

2 Theory

2.1 Genomic Structure

Bacteria usually have one circular chromosome which holds most of the genes. They can also have smaller DNA fragments, called plasmids. At the origin of replication (*oriC*), they have specific DNA motifs and genes initiating genome replication which most often is bidirectional. Bidirectional replication is when two replication forks perform the replication simultaneously in opposite directions from the *oriC*, and the replication is ended at the opposite side of the chromosome, or plasmid, at the terminus of replication (*ter*) (Deng *et al.* 2016). The bacteria genome is mostly made up of genes, which means that the genome only has a small proportion of non-coding parts (Neville & O'Toole 2014). Furthermore, genes that are functionally related to each other in bacteria can be placed after each other in the genome in operons (Neville & O'Toole 2014). The genes included in an operon are transcribed and translated together (Neville & O'Toole 2014). Genes in a species can be classified into different gene families, and a gene family consists of genes related to each other by duplication from a common ancestor (Lal *et al.* 2020).

2.2 Pan-genome, Core Genome, and Accessory Genome

Tettelin *et al.* (2005) were the ones that coined the expression "pan-genome" to describe the total, non-redundant, genetic material of a species, including all parts of the genomes found in all different strains. Since then, the pan-genome has become a well-known phrase, and the interest to study the pan-genome has grown through the years. The pan-genome can be divided into the core genome, the accessory genome (sometimes called dispensable genome), and the unique genes (Medini *et al.* 2005). The strict definition of the core genome is that it includes all genes that are found in all strains for a species, but sometimes a less strict definition can be used and the core genome can e.g. be defined as the genes found in at least 95 % of the strains (Medini *et al.* 2005). The less strict definition can be useful when analysing larger datasets because the risk of not finding genes that are in the genomes is increasing with a larger dataset. Genes included in the core genome are often essential for the cell's survival and among others, the core gene includes genes for regulatory functions, cell growth, and housekeeping functions. The accessory genome includes the genes not present in all strains (or found in fewer strains than the limit for the core genome), but found in at least two genomes (Medini *et al.* 2005). The unique genes are the set of genes found in only one strain (Medini *et al.* 2005). However, the definition of a unique gene is dependent on the size of the dataset that is analysed, how it was sampled, and how a strain is defined. If more strains were added to the analysis, there is the chance that many of the unique genes would be found in more than one strain and instead be part of the accessory genome.

The pan-genome for a species can be said to be closed or open. When a pan-genome is closed, it means that when new genomes are added to the analysis, the number of new genes are decreasing towards zero. This means the species pan-genome might be fully characterized with the genomes included in the analysis (Medini *et al.* 2005). If a pan-genome is open, the number of new genes added for each new genome will stabilize around a certain number of genes (Medini *et al.* 2005). After this point, on average the same number of genes will be added to the pan-genome with each added genome (Medini *et al.* 2005). A way to estimate the openness of a pan-genome and its growth rate is to use Heaps' law (Tettelin *et al.* 2008, Park *et al.* 2019), which is defined as:

$$n = \kappa N^{\gamma} \tag{1}$$

Where n is defined as the size of the pan-genome, κ and γ are parameters for fitting, and N is the number of genomes included in the pan-genome (Park *et al.* 2019). The pan-genome is estimated to be open when $\gamma > 0$, and closed when $\gamma < 0$ (Tettelin *et al.* 2008, Park *et al.* 2019). Another way to estimate the openness is to use a power law model that finds how the rate of new genes added to the pan-genome is decreasing (Tettelin *et al.* 2008, Park *et al.* 2019), the power law model is defined as:

$$\Delta n = \kappa N^{(-\alpha)} \tag{2}$$

Where Δn is defined as the number of genes added, κ and α are parameters for fitting, and N is the number of genomes included in the pan-genome (Park *et al.* 2019). A pan-genome is estimated to be open when $\alpha \leq 1$, and closed when $\alpha > 1$ (Tettelin *et al.* 2008). Of the species studied in this thesis, *L. monocytogenes* has previously been found to have a closed pan-genome (Halachev *et al.* 2011). *C. jejuni*, *E. coli*, *S. enterica*, and *S. pneumoniae* have previously been found to have open pan-genomes (Halachev *et al.* 2011, Park *et al.* 2019).

Up until today, a lot of approaches and software have been made to find a species pan-genome as well as dividing it up into core genome and accessory genome (Vernikos 2020). Some of the software compares the genomic sequences by whole genome alignment and thereby divides the genome into the core- and accessory genome (Ozer *et al.* 2014). A comparison by whole genome alignment is a good option to make sure not to miss any non-coding regions in the genomes, however, these methods are very time consuming, and is therefore only suitable to use when working with small datasets (Ozer *et al.* 2014). To reduce the running time and optimize the usage of Random Access Memory (RAM), other methods have been developed that instead use clustering of genes to divide them up into core- and accessory genes (Page *et al.* 2015, Chaudhari *et al.* 2016).

2.3 Bacterial Pan Genome Analysis tool

Bacterial Pan Genome Analysis tool (BPGA) is a tool for pan-genome analysis which divides protein-coding genes into core-, accessory-, and unique genes by clustering similar sequences together. The tool takes genomes as input in different formats, e.g. GenBank files or FASTA files. After entering the input files, BPGA prepares the data for clustering. The next step is the clustering, where the protein sequences from the genes are clustered together based on their similarity. For repeated genes, BPGA only considers the first time the gene occurs in the analysis, and the gene is only clustered once. The user has the option to choose from three different clustering methods: USEARCH, CD-HIT, and OrthoMCL. USEARCH is said to be the fastest clustering method and is the default method for the tool. When proteins have been clustered together, one cluster represents one gene family and the proteins in the cluster can be interpreted as one protein representing its corresponding gene. From the clustering, BPGA divides the genes into core-, accessory-, and unique genes depending on how many genomes the different clusters (genes) are found in. The genes found in all genomes analysed are classified as core genes, the ones not found in all but at least in two genomes are classified as accessory genes, and the ones found in only one genome are classified as unique genes. From this, BPGA calculates the pan-genome as well as the core-genome. The openness of the pan-genome is estimated with Heaps' law, and the program gives the fitting parameters κ and γ. In addition to this, BPGA has several downstream analyses that are optional to use. One of these is the option to map the classified genes towards Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and to the Clusters of Orthologous Groups (COG) categories. BPGA has been validated to work for 1,000 genomes where it was seen that it had a lower run time and used less RAM than other clustering tools for pan-genome analysis. The lower running time and the decrease in RAM usage give BPGA an advantage over other methods when analysing large datasets. (Chaudhari *et al.* 2016)

2.4 KEGG Pathways and COG Categories

KEGG is a project initiated in 1995, it was created to connect genetic information to functional information (Kanehisa & Goto 2000). Today, KEGG is a resource that consists of 18 different databases and enables the understanding of biological systems such as cells, organisms, and ecosystems at the molecular level (Kanehisa *et al.* 2021). The databases are divided up into four categories: systems information, genomic information, chemical information, and health information (Kanehisa *et al.* 2021). KEGG pathways is one of the eighteen databases and is found in the systems information category (Kanehisa *et al.* 2021). This database consists of manually drawn maps of KEGG pathways, which represents the knowledge for molecular interactions, reactions, and relations (KEGG 2021). The pathways are divided into seven categories: metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, human diseases, and drug development (KEGG 2021). The first six categories are used by BPGA when it maps classified genes toward KEGG pathways. All categories in KEGG pathways are further divided into more specific categories which can be seen in Appendix A for the categories used by BPGA.

COG is another database created in 1997 that is used to predict the function of proteins (Tatusov 2000, COG 2021). The COG database added functionality from proteins that have been experimentally characterized (Tatusov 2000). Thereafter, functional predictions have been added to proteins coming from poorly studied organisms by finding their orthologous relationship to the proteins that were characterized experimentally (Tatusov 2000). Today, the database holds complete genomes from 1,187 bacteria, and the proteins are divided into 26 different categories based on their functionality (COG 2021). When BPGA map proteins towards COG, it maps against the COG database released in 2003, which had 20 COG categories (Chaudhari *et al.* 2016). The COG categories are summed up into four larger categories for COG in the BPGA analysis: cellular process and signalling, information storage and processing, metabolism, and poorly characterized.

3 Materials and Methods

3.1 Computational Resources

During this project, the scripts used were written in Python 3.8.5 (Van Rossum & Drake 2009). The pan-genome analyses were performed with BPGA 1.3.0 (Chaudhari *et al.* 2016). BLAST 2.9 was used to blast core genes towards reference genomes (Altschul *et al.* 1990). All computations ran locally on a computer having one processor with eight cores, each core having two threads. The computer had a RAM of 156 gigabytes.

3.2 Data

In this study, I used data that I got from my supervisor. The data was originally obtained from the National Center for Biotechnology Information (NCBI) genome database that contained sequenced genomes presented as GenBank files. The genomes came from surveillance projects from different institutions all over the world. All genomes in the database were already assembled and annotated. In the database, there were 49 different bacterial species present, and the six species with the highest number of genomes present in the database were chosen for this project: *C. coli* (12,904 genomes), *C. jejuni* (34,002 genomes), *E. coli* (34,222 genomes), *L. monocytogenes* (29,857 genomes), *S. enterica* (243,401 genomes), and *S. pneumoniae* (13,437 genomes). For each of the chosen species, smaller datasets of 10, 100, 1500, 2000, and 2500 genomes were constructed. The genomes in the smaller datasets were randomly chosen from the complete datasets of each species. The GenBank files were put in separate folders for each species and each size of the dataset.

3.3 Pan-genome Construction

BPGA was used for the pan-genome analysis of the six species. For each run with BPGA, the GenBank files in the datasets to be analysed were unzipped and entered as input files. Through all steps of BPGA, it was running with default settings and parameters. BPGA performed a preparation of the data and then the clustering step was initiated. USEARCH was used for the clustering step. USEARCH performs a sequence alignment on the protein sequences derived from the genes. Then it clusters the protein-coding genes together based on their proteins similarity. After the clustering, BPGA divided the genes into core-, accessory-, and unique genes. Furthermore, it calculated the pan-genome and core genome, and used Heaps' law to estimate the openness of the pan-genome. BPGA was also used to map the classified genes to COG categories and KEGG pathways.

The datasets were analysed by BPGA one at a time. First, the smallest datasets having ten genomes were run with BPGA for each species. Then the number of genomes to be analysed for each species were scaled up by increasing the size of the datasets. This was done for all species until BPGA could no longer handle the number of genes entered for analysis. The reason it could not analyse a higher number of genomes were not that the computer's capacity of RAM was maximized, it had to do with the scripts used inside of BPGA.

For the resulting datasets analysed with BPGA, the mean number of genes in each species genomes were calculated and rounded to be an integer value. A relative core genome size was calculated for the species by dividing the number of genes classified as core genes with the mean number of genes in the genomes for the species. The relative core genome sizes can be expected to have approximately the same size if more genomes were added to the analysis. A relative accessory genome size was calculated too, where the sum of accessory-, and unique genes were divided by the mean number of genes in the genomes. This shows how many accessory-, and unique genes the species have in relation to their genome sizes in the datasets analysed in this project. However, I would like to emphasize that this measurement cannot be predicted for other datasets like the relative core genome size can. The percentage given for the relative accessory genome size is a measurement for how many accessory-, and unique genes were found in the specific datasets of genomes analysed here. However, if more genomes were added to the analysis, the percentages would most probably change for all species since other accessory-, and unique genes will be found in the added genomes.

3.4 Frequency of Accessory Genes

Text files were outputted by BPGA with information about the genes that were classified in the pan-genome analysis. In each text file, all the protein sequences from the classified genes were presented together with information about which genome the protein-coding gene came from, which cluster the protein had been clustered to, the protein accession number, and the classification (core, accessory, or unique). To analyse the gene frequency of all genes in the species, a Python3 script (Van Rossum & Drake 2009) was constructed. The text files were scanned through to find which cluster ids were present in the file. Then the number of proteins in each cluster were calculated. The number of proteins in each cluster represented how many genomes each protein, or gene, was found in.

3.5 Combining Gene Annotation Data with BPGA Classifications

To be able to analyse the conservation of the core genes synteny, the structural annotation of the genes from the genomes included in the pan-genome analysis had to be combined with the cluster ids and classifications from BPGA. The information was gathered and saved as a table in a CSV file, forming a gene database for each species. The GenBank files and the text files from BPGA were mined for information about the genome and its genes by constructing a Python3 script (Van Rossum & Drake 2009) and using the BioPython package SeqIO (Cock *et al.* 2009). For each gene in each genome and GenBank file, the following information was gathered and added to a table: genome accession number, contig accession number, contig number, contig size, gene number in the contig, assembly method (de novo or reference guided), gene type (protein-coding, rRNA, tRNA, etc.), locus tag (gene identifier), gene annotation, protein accession number, pseudogene or not, coordinates in the genome (together with information if the gene was on the forward or reverse strand). Some genes had several annotations, these were saved together in one field in the table and separated by a semicolon. A unique id was created for each gene by combining the gene accession number with the contig number and the gene number. Two fields were added in the table with information from the text files from BPGA: cluster id and the classification of the genes (core, accessory, or unique). The information from the text file was found by matching the protein accession numbers from the GenBank file with the protein accession numbers in the BPGA files. Since BPGA only classified protein-coding genes, not all genes in the table had a cluster id and a classification. For genes not classified by BPGA, no cluster id was stated as *'No'* and no classification was stated as *'0'* in the table. The table was saved as a CSV file for each species.

3.6 Analysis of Core Gene Synteny

The gene database presented in section 3.5 was used in an analysis to look at the conservation of synteny in the core genes for all species one at a time. A Python3 script (Van Rossum & Drake 2009) was written and used to carry out the analysis. BPGA did not classify tRNA and rRNA coding genes since it only works with protein-coding genes. Therefore, in this analysis, tRNA and rRNA coding genes were classified as core genes and were given a cluster id each. For each gene classified as a core gene, the next downstream core gene was identified by using the gene database. By this, the order of the core genes (represented by their corresponding cluster id) could be compared between the genomes in a species. In this analysis, it did not matter if a core gene sat directly after another core gene in the genome or if there were other genes in between them (accessory-, unique-, or not classified genes). A schematic image exemplifying this is shown in Figure 1. For each core gene, the total number of times it was followed by all other core genes was calculated. The core gene that most often followed a core gene was chosen, and the two genes together formed the most common core gene pair. The number of times the most common core gene pair was present was divided by the total number of times the upstream core gene in the core gene pair was present in the genomes and having a downstream core gene connected to it. It was divided by this number since some genes are located at the end of a contig, hence not having a downstream core gene following it in this genome. This gave a percentage for how often the most frequent following core gene did follow the core gene. If a core gene were given a score of 100 %, it meant that the core gene was always followed by the same downstream core gene and the core gene pair had full conservation regarding the synteny in the genomes that were included in the analysis.

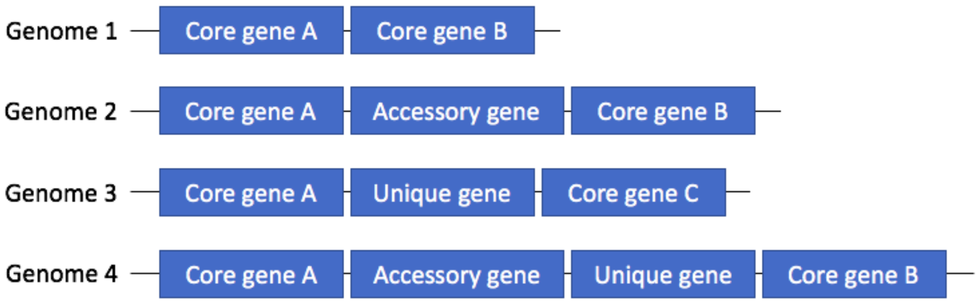


Figure 1. Schematic image showing how the analysis for core genes following each other in the genomes were working. In genome 1, 2, and 4 core gene A is followed by core gene B even though there are other non-core genes between core gene A and core gene B in genome 2 and 4. In genome 3, core gene A is instead followed by core gene C.

For each conserved core gene pair, the gene database was used again to count how many times the two core genes sat directly after each other in the genome. It was also calculated how often the first core gene in the core gene pair was instead directly followed by an accessory gene, and how often it was followed by a not classified gene. In this analysis, both accessory genes and unique genes were counted as accessory genes. The calculated numbers were used to calculate a percentage showing how often the upstream core gene in the core gene pair was followed by a core-, accessory-, or not classified gene.

The next step was to analyse how large the accessory segments sitting between the core genes in each core gene pair were. The size of the accessory segments was defined as the number of accessory genes in the segments. For each core gene pair where accessory genes sometimes or always were found between the core genes, the gene database was used to count the size of the accessory segments. Maximum length, minimum length, median length, first quartile, and third quartile was calculated for the accessory segments occurring between each core gene pair.

3.7 Comparison Against Reference Genome

The core genes were compared with a reference genome for all species, by writing a Python3 script (Van Rossum & Drake 2009) and using BLAST (Altschul *et al.* 1990), to see how the genes classified by BPGA were distributed in a complete reference genome. One reference genome for each species was downloaded from NCBI. The name of the reference genomes and their respective accession number can be found in Table 1. For each species, the core genes were divided up into three categories:

1. Core genes always or almost always followed directly by another core gene (followed by a core gene in at least 95 % of the cases).
2. Core genes sometimes followed directly by core genes (followed by a core gene in 5 % - 95 % of the cases).
3. Core genes never or seldom followed by core genes (followed by a core gene in up to 5 % of the cases).

The text files from BPGA having the protein sequences were used to take out one representative sequence for each protein-coding core gene. In this step, the tRNA and rRNA coding genes were no longer present in the analysis as core genes, since they were not classified by BPGA. For all three categories of core genes, the cluster id for each core gene was used to take out all protein sequences representing a gene in a cluster. One of the protein sequences were randomly chosen to be the representative sequence for each gene. The three categories of core genes were then blasted, with tblastn, towards the reference genome one at a time for each species. From the BLAST output, only the best hits were filtered out for each gene, this was done by choosing the hit that had the best score. The start positions in the reference genome were saved for the best hits for all three categories of core genes.

Table 1. Reference genomes for each species from NCBI and their accession numbers.

Species	Reference genome	Accession number for reference genome
Campylobacter coli	Campylobacter coli strain aerotolerant OR12, complete genome	NZ_CP019977.1
Campylobacter jejuni	Campylobacter jejuni subsp. jejuni NCTC 11168 = ATCC 700819 chromosome, complete genome	NC_002163.1
Escherichia coli	Escherichia coli str. K-12 substr. MG1655, complete genome	NC_000913.3
Listeria monocytogenes	Listeria monocytogenes EGD-e chromosome, complete genome	NC_003210.1
Salmonella enterica	Salmonella enterica subsp. enterica serovar Typhimurium str. LT2, complete genome	NC_003197.2
Streptococcus pneumoniae	Streptococcus pneumoniae strain NCTC7465 chromosome 1, complete sequence	NZ_LN831051.1

4 Results

4.1 Construction of Pan-genomes

Pan-genome analyses were made for each species by using BPGA. I could successfully analyse 1,000 genomes for *E. coli* and *S. enterica*, and 2,000 genomes for *C. coli*, *C. jejuni*, *L. monocytogenes*, and *S. pneumoniae*. When trying to analyse a larger dataset than this, the program gave errors indicating that the internal scripts in BPGA performing the analysis in the program could not handle the data size. However, the errors did not appear to be related to lack of RAM.

4.1.1 Pan-genome Construction with BPGA

BPGA was used to classify the protein-coding genes for each species into core-, accessory-, and unique genes, to perform a pan-genome analysis, and to estimate the openness of the pan-genomes. The result from this is shown in Table 2 and Table 3. As seen in Table 2, *E. coli* and *S. enterica* had the largest genome sizes, both having around 5,000 genes on average in the genomes analysed. They also had the highest number of accessory-, and unique genes relative to the genome size. The relative core genome size represented how many genes in the mean genome size were classified as core genes, and was presented as a percentage. *C. jejuni*, *L. monocytogenes*, and *S. enterica* had the highest relative core genome size with a percentage of around 40 %. *C. coli* and *E. coli* had the lowest relative core genome size around 40 %. However, *C. coli's* relative core genome size was probably underestimated because of an artefact where a wrongly annotated genome was part of the analysis for *C. coli*. The artefact is described in more detail in the next section (4.1.2). *E. coli* and *S. enterica* had the largest pan-genomes and *C. jejuni* had the smallest pan-genome size.

4.1.2 Openness of the Pan-genomes

During the pan-genome analysis with BPGA, genes were cumulatively added to the pan-genome and taken away from the core genome as new genomes were added iteratively to the comparison. Together with the total number of gene families in the pan-genome and core genome, it produced estimated curves representing the cumulative growth of the pan-genome and the cumulative shrinkage of the core genome with the growing number of genomes. A visualisation of this is shown in Figure 2. For most of the species, except *E. coli*, the pan-genome had flattened out considerably at the end of the curves. During the pan-genome analysis, BPGA used curve fitting with Heaps' law to estimate the openness of the pan-genomes, shown in Table 3. All species had a $\gamma > 0$, which means they were estimated to have open pan-genomes. *E. coli* and *S. enterica* had the largest γ value, just over 0.2, indicating a higher openness, while *C. jejuni* had the lowest γ value.

The cumulative growth of gene families for the pan-genome and the core genome for *C. coli* shows that one genome, around gene number 1,250, contributed with many new gene families and the core genome were thereby decreased a lot for the same genome. Because of this, I looked up which genome contributed to this and found out that this genome contributed with 955 new genes to the pan-genome which is considerably more than expected. A manually random chosen part of the genome was chosen from the GenBank file for this genome. The genome segment was run with BLAST (Altschul *et al.* 1990) against the GenBank NR database. This gave a 100 % query cover match for *Campylobacter fetus (C. fetus)*, which suggests that this genome most probably has been wrongly annotated and is a *C. fetus* instead of a *C. coli*. Because of the time limit for this project, there was no time to redo the analysis without the wrongly annotated genome after it was discovered. Furthermore, *C. jejuni* also had two spots in the pan-genome curve where a larger number of new genes were added, however, these were much smaller than the one found in *C. coli*.

4.1.3 Frequency of Classified Genes

For all genes classified by BPGA during the pan-genome analysis, it was calculated how many genomes each gene, represented by their cluster id, was found in. This was used to visualise the gene frequency distribution, shown in Figure 3. To the left of the red vertical line in the plots are the core genes which were found in all genomes. To the right of the red line, the accessory and the unique genes can be seen. For all species, there was a group of accessory genes that were present in more than 95 % of the genomes but not in all, these are shown between the red line and the black vertical line. In a less stringent core genome definition, these genes would be considered part of the core genome, hence the genes can be called low stringency core genes. All species except *C. coli* had more genes classified as core genes than they had low stringency core genes. However, the higher amount of low stringency core genes for *C. coli* is probably because of the artefact in the analysis that a wrongly annotated genome of the species *C. fetus* were a part of the *C. coli* pan-genome analysis. Approximately half the genes were found in only one or just a few genomes for all species.

4.1.4 Functional Classification of Core-, Accessory-, and Unique Genes

BPGA was used to map the core-, accessory-, and unique genes to COG categories and KEGG pathways to get insight into which biological processes the genes takes part in. The gene distribution among the major COG categories is found in Figure 4. Some genes were classified as poorly characterized, which means they could not be given functionality from the COG categories. For all species, the largest portion of the core genes was found to be connected to metabolism. The accessory-, and unique genes instead had their largest portion of genes connected to information storage and processing for *E. coli*, *L. monocytogenes*, *S. enterica*, and *S. pneumoniae*. The species *C. coli* and *C. jejuni* had most of their accessory-, and unique genes connected to metabolism. All species had a higher portion of their accessory-, and unique genes than their core genes classified as poorly characterized. A more detailed visualisation of the COG distribution with more specific categories can be seen in Appendix B.

The distribution of the genes mapped to the major KEGG pathways is seen in Figure 5. Similarly to what was seen in the COG analysis, the core genes had a higher percentage of their genes connected to metabolism pathways than the accessory-, and unique genes for all species. This with an exception for *C. jejuni*, where the accessory genes had a slightly higher percentage connected to metabolism. Furthermore, the core genes had a higher proportion of its genes found in the pathways for genetic information processing than the accessory-, and unique genes had for all species. Two categories of pathways showed a higher representation of accessory-, and unique genes than core genes: environmental information processing, and human diseases. A more detailed visualisation of the KEGG distribution with more specific categories can be seen in Appendix C.

Table 2. Summary of the number of genomes and genes analysed, together with the number of genes classified by BPGA, and the estimated pan-genome sizes.

			Number of genes classified by BPGA					
Species	Number of genomes in pan-genome analysis	Mean number of genes in genomes	Core	Accessory	Unique	Relative core genome size [%]	Relative accessory genome size [%]	Pan-genome size [number of genes]
Campylobacter coli	2,000	1,890	554	3,457	1,523	29	263	5,534
Campylobacter jejuni	2,000	1,844	736	3,069	1,077	40	225	4,882
Escherichia coli	1,000	5,317	1,662	16,026	4,803	31	392	22,491
Listeria monocytogenes	2,000	3,100	1,330	5,826	2,321	43	262	9,477
Salmonella enterica	1,000	4,775	1,977	13,247	5,411	41	391	20,635
Streptococcus pneumoniae	2,000	2,134	775	4145	965	36	239	5,885

Table 3. Results from pan-genome analysis. Fitting parameters (***κ***, ***γ***) from Heaps' law, open or closed pan-genome.

Species	κ	γ	Open / closed pan-genome
Campylobacter coli	1206.02	0.191656	Open
Campylobacter jejuni	1747.07	0.126668	Open
Escherichia coli	4729.21	0.214124	Open
Listeria monocytogenes	2568.04	0.174854	Open
Salmonella enterica	3984.4	0.24002	Open
Streptococcus pneumoniae	2076.87	0.137829	Open

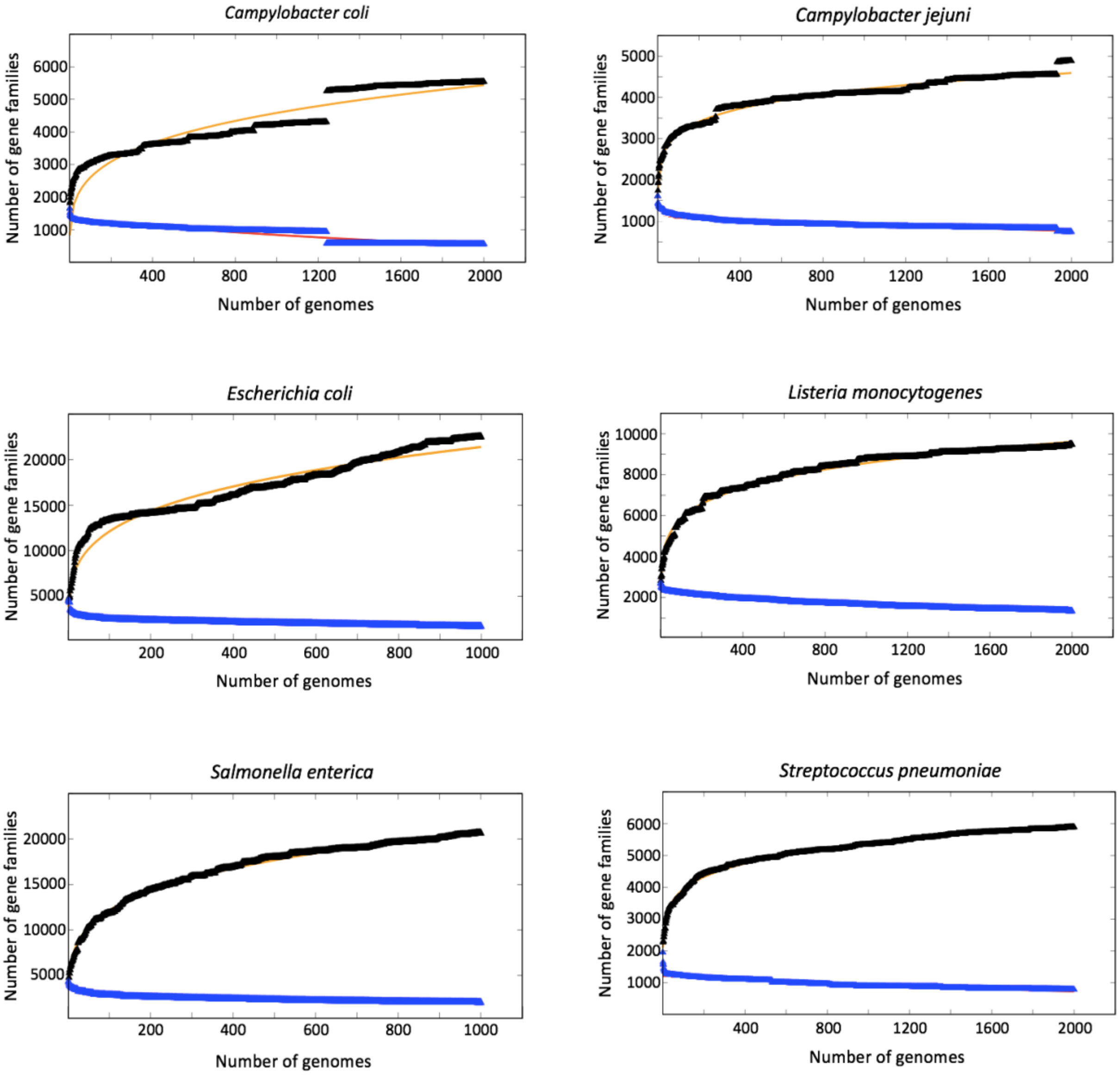
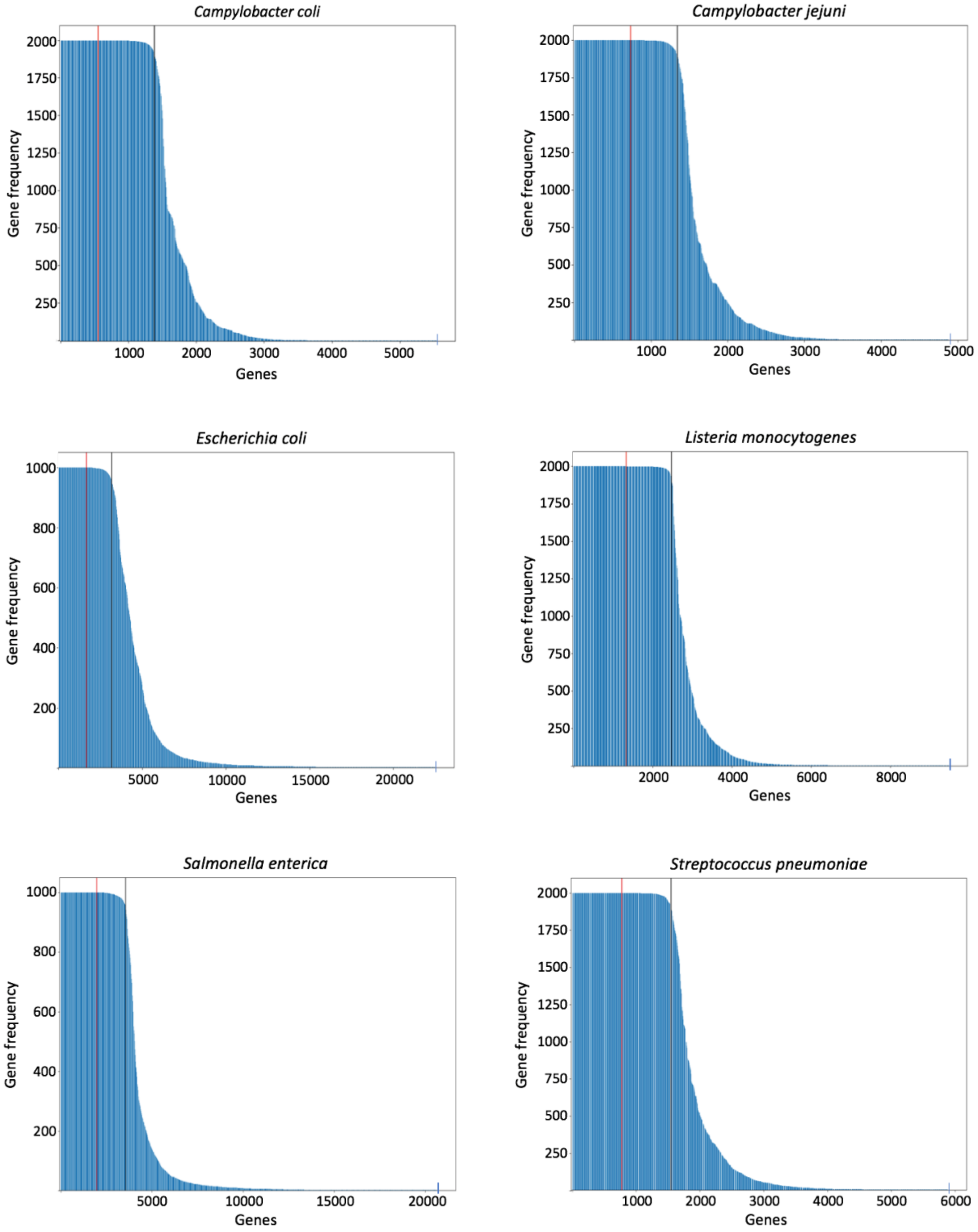


Figure 2. Cumulative growth curves (orange) and core genome curves (red) based on curve fitting using Heaps' law, together with the cumulative total number of gene families for pan-genome (black) and core genome (blue). The orange and red lines are for some species totally or partly hidden behind the black and blue lines.



Campylobacter coli

Campylobacter jejuni

Escherichia coli

Listeria monocytogenes

Salmonella enterica

Streptococcus pneumoniae

Figure 3. Gene frequency distribution for genes classified by BPGA. The red line indicates the transition from core genes to accessory genes. The field between the red and the black line shows the accessory genes found in at least 95 % of the genomes. A larger thick on the x-axis has been added at the end of the curves, indicating where the last unique gene is found and how many genes are seen in the plots.

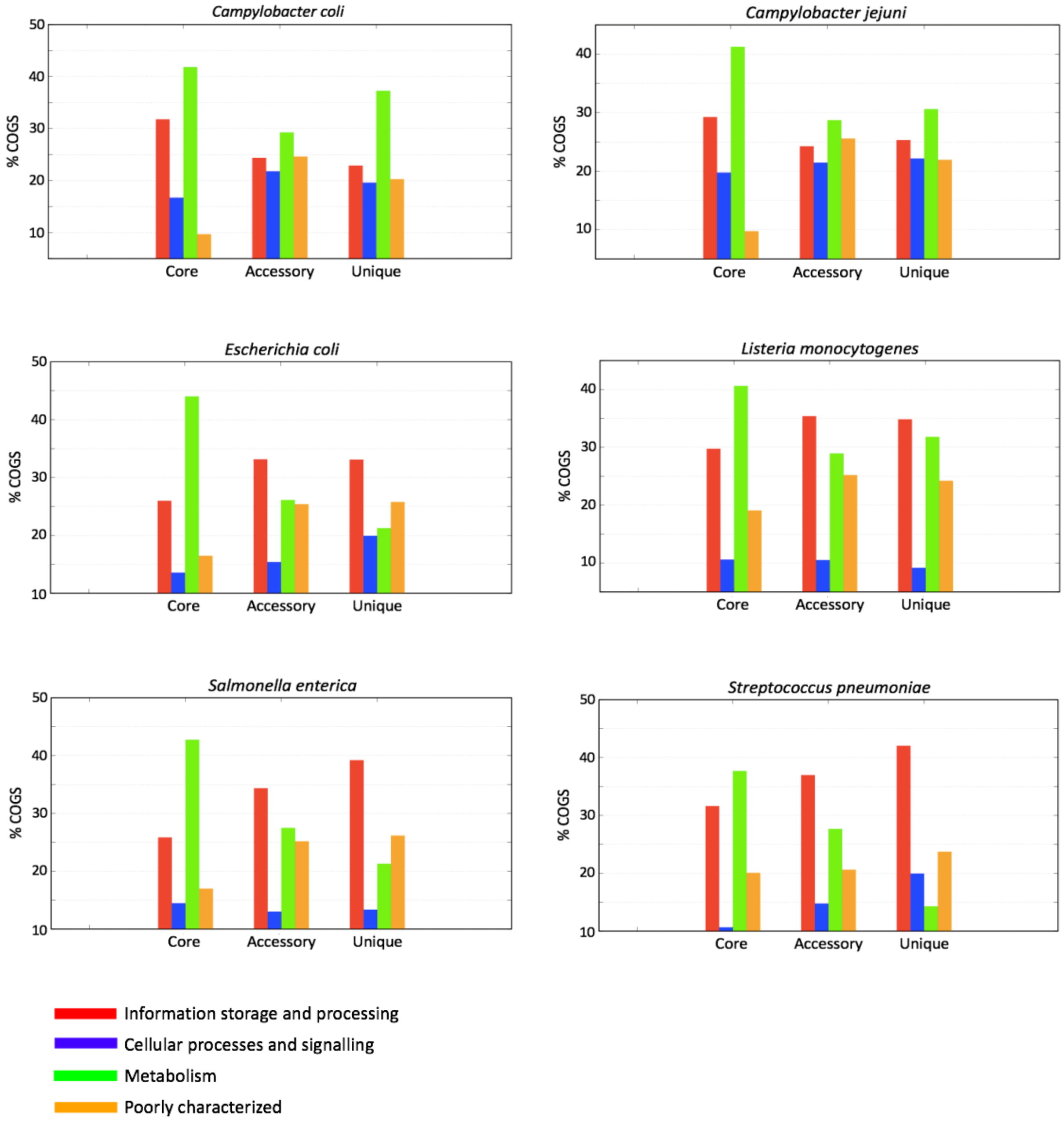


Figure 4. COG distribution of genes classified by BPGA.

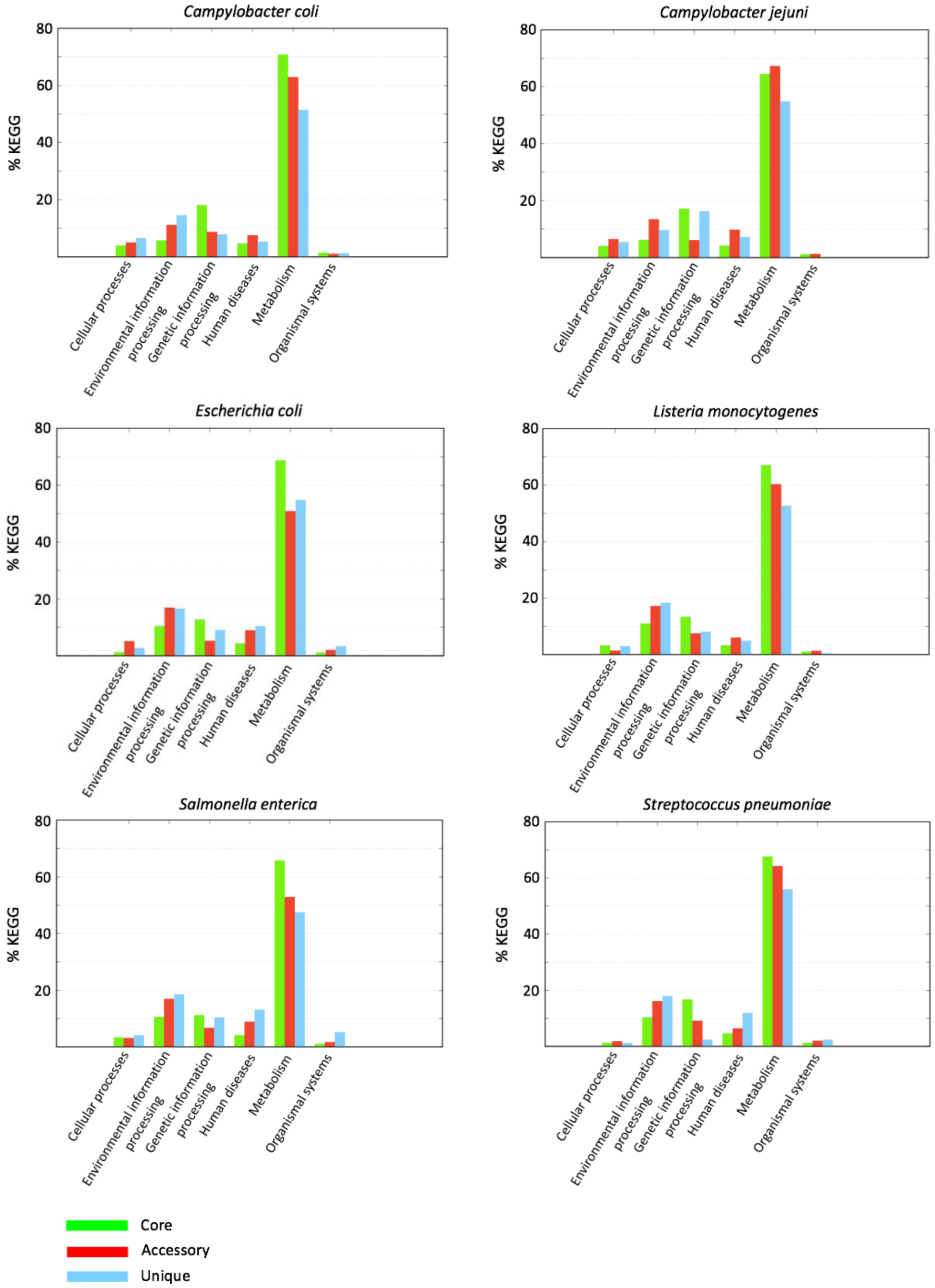


Figure 5. KEGG distribution of genes classified by BPGA.

4.2 Analysis of Core Gene Synteny and Distribution of Core Genes

To get a better understanding of the core genes' properties and to investigate if there were sites in the genomes more prone to take in new genes, the core gene synteny was analysed. The core genes were also compared to one reference genome for each species to visualise their distribution along that genome.

4.2.1 Conservation of Core Gene Synteny

For all core genes in the genomes of each species, an analysis was made to find out which downstream core gene most often followed the upstream core gene, the two core genes forming a core gene pair. To be able to compare the gene order between genomes, the order of the assigned cluster ids for the core genes were used. For each core gene pair, its frequency was calculated as a percentage of the total number of times the core gene pair was present in the genomes of a species. The percentage can be interpreted as a measurement for how conserved different core genes were regarding their synteny, and the result from this is shown in Figure 6. *S. pneumoniae* stood out from the other species because it was the only species not having any conserved genes with a percentage over 90 %, while all the other species had some genes at 100 % or almost 100 %. Furthermore, *S. pneumoniae* only had very few genes over 80 % and almost half of its genes had a conservation percentage below 60 %. For the other species, a clear majority of the genes had a conservation percentage above 60 %. They also had a considerable number of genes above 80 %. *C. jejuni*, *E. coli*, and *S. enterica* showed the largest number of genes having more than 80 % conservation.

For each conserved core gene pair presented in Figure 6, it was analysed how often the conserved core genes directly followed each other and how often accessory genes or not classified genes sat between the core genes. During the analysis, both accessory genes and unique genes were counted as accessory genes. The result from the analysis is shown in Figure 7. The percentage is a measurement for how often the two conserved core genes sat directly after each other, and how often accessory-, or not classified genes sat between the core genes. All species had a considerable number of core gene pairs where no accessory genes or not classified genes appeared between them. They also had a smaller number of core genes where accessory genes or not classified genes always sat between the core genes. However, there was an exception for *C. coli* where the number of core genes always having other genes between them and the number of core genes never having other genes between were approximately the same amount. Because of a technical artefact, *C. coli* had core genes that most probably were misclassified as accessory genes, which may influence this result. For all species, when looking at the percentage that the core genes sat directly after each other, it decreased at a steady rate from 100 % to 0 % for almost all species. However, for *S. pneumoniae* it did not decrease in the same way, it had very few genes where core percentage was between 0-30 % and 70-100 %, which made the shape of the plot different from the other species.

4.2.2 Size of Accessory Segments

The length of the accessory segments (number of accessory genes) inserted between conserved core gene pairs was analysed. For each species, the median, maximum, and minimum length of the accessory segments are presented in Figure 8. The median, first quartile, and third quartile of the accessory segments lengths are shown in Figure 9. The maximum length of accessory segments showed that all species did take in quite large accessory segments between core genes, even when the median length goes down. There was no obvious association between the maximum values and the median values. *E. coli* and *S. enterica* were the species with the highest number of large accessory elements when looking at the maximum lengths. Both species had several genes that had taken in accessory segments with a size over 100 accessory genes. For the other species, the maximum length of accessory segments reached up to 70-80 accessory genes, except for only one core gene pair that had an accessory segment with over 100 genes for both *C. jejuni* and *L. monocytogenes*. There was a bit more association between the minimum length and the median length for all species, meaning there were probably a higher amount of short accessory segments than large ones sitting between the core gene pairs. Some of the accessory segments with a larger median length also had a larger minimum value, which indicates that these spots in the genomes always take in larger accessory segments. However, this was not true for all genes with a higher median value. When looking at the first quartile and third quartile values, the outliers in the datasets are ignored, seen in Figure 9. From this, it is possible to see how many core gene pairs tend to take in larger accessory segments more frequently than others. *C. jejuni*, *E. coli*, and *L. monocytogenes* all had one gene each where the third quartile value was considerably high even though the median value had gone down. This indicates that these genes had high amounts of both small and large accessory segments, large accessory segments sat between the core genes several times for those core gene pairs.

4.2.3 Comparison Against Reference Genome

All conserved core genes were split up into three categories for each species:

1. Core genes always or almost always followed directly by another core gene (in Figure 7: core > 95 %).
2. Core genes sometimes followed directly by core genes (in Figure 7: 95 % > core > 5 %).
3. Core genes never or seldom followed by core genes (in Figure 7: core < 5 %).

The three categories of core genes were aligned with BLAST to a reference genome to get insight into the genes' placement. The cumulative gene numbers, represented as percentages, were plotted against the genes start positions in the reference genomes. This gave an overview of the distribution of the genes from the three categories over the reference genomes. The resulting plots are shown in Figure 10. For some species and some categories of core genes, the genes were evenly placed out in the reference genome. This is shown where the genes were laying up on a straight line with a constant slope. *E. coli* is an example where all three categories of genes were quite evenly placed in the reference genome. When the genes had a steeper slope in a part of the genome, it indicated a higher gene density in that area of the genome, i.e. more core genes were present in this region of the reference genome. Steeper slopes were found in all species but were clearer for some of them. *C. coli* had several regions for all categories of genes where the gene density was high and the slope was steep, sometimes the slope was almost completely vertical. *C. jejuni* also showed this tendency for all categories. Both *C. coli* and *C. jejuni* had a high gene density at the end of the genome for the core genes in category 1 (core genes always followed by a core gene). *L. monocytogenes* had regions with high gene density for the core genes always followed by a core gene and for the core genes never followed by other core genes. For *S. pneumoniae*, the regions with the highest gene density were found for genes in category 1 (core genes always followed by a core gene). However, the other categories also had regions with slightly higher gene density. Generally, when the different categories of genes do not lay on a line together, but their lines get separated, this means that they had a high gene density in different parts of the genome. Furthermore, *C. coli* had a large gap in the reference genome where no genes were found, the other species had smaller gaps in some places but none of them was as large as for *C. coli*.

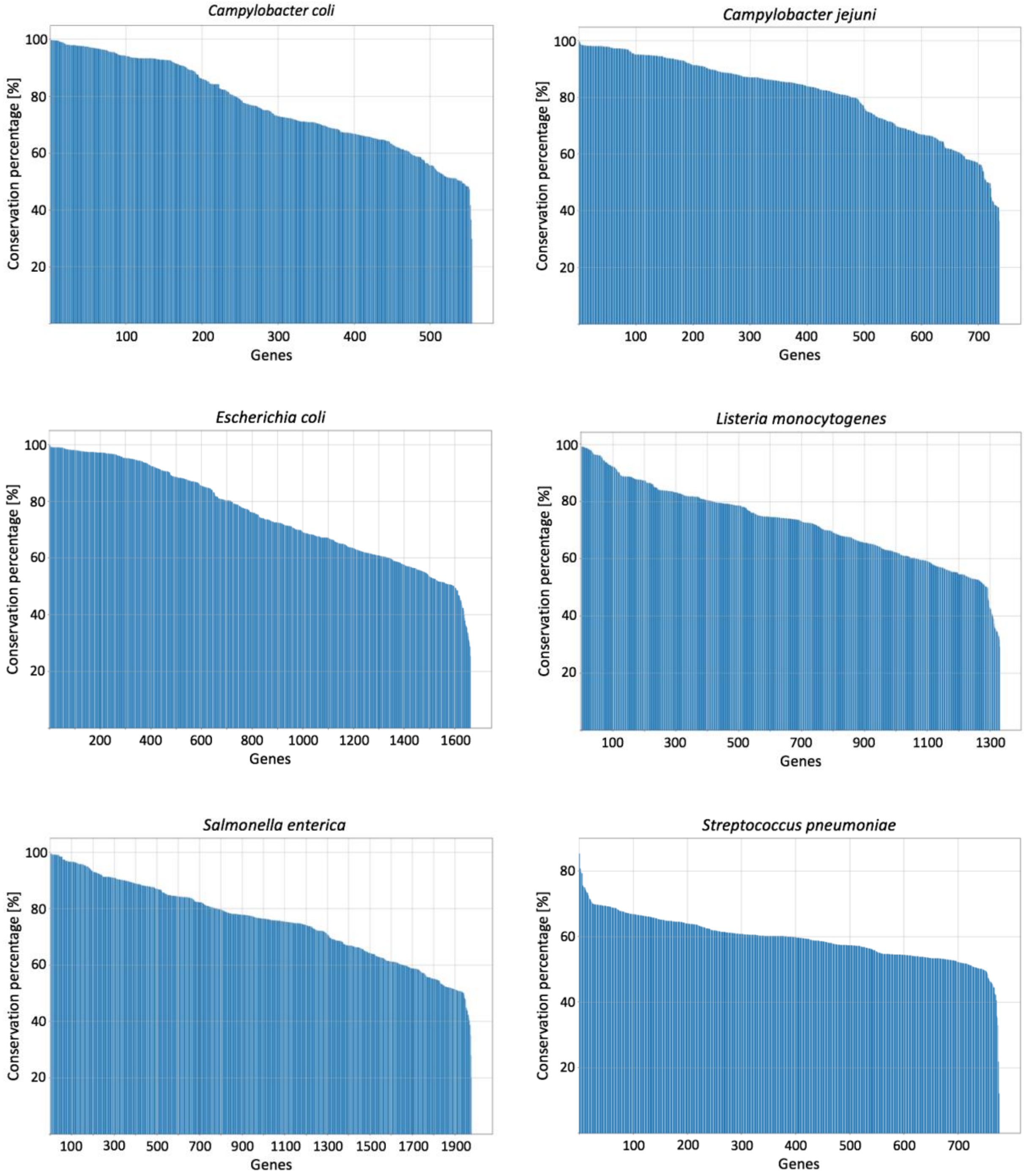


Figure 6. Conservation percentage for the core gene synteny. For each core gene, the percentage represents the frequency for how often the gene was followed by the downstream core gene that most often sat after the gene in the genomes.

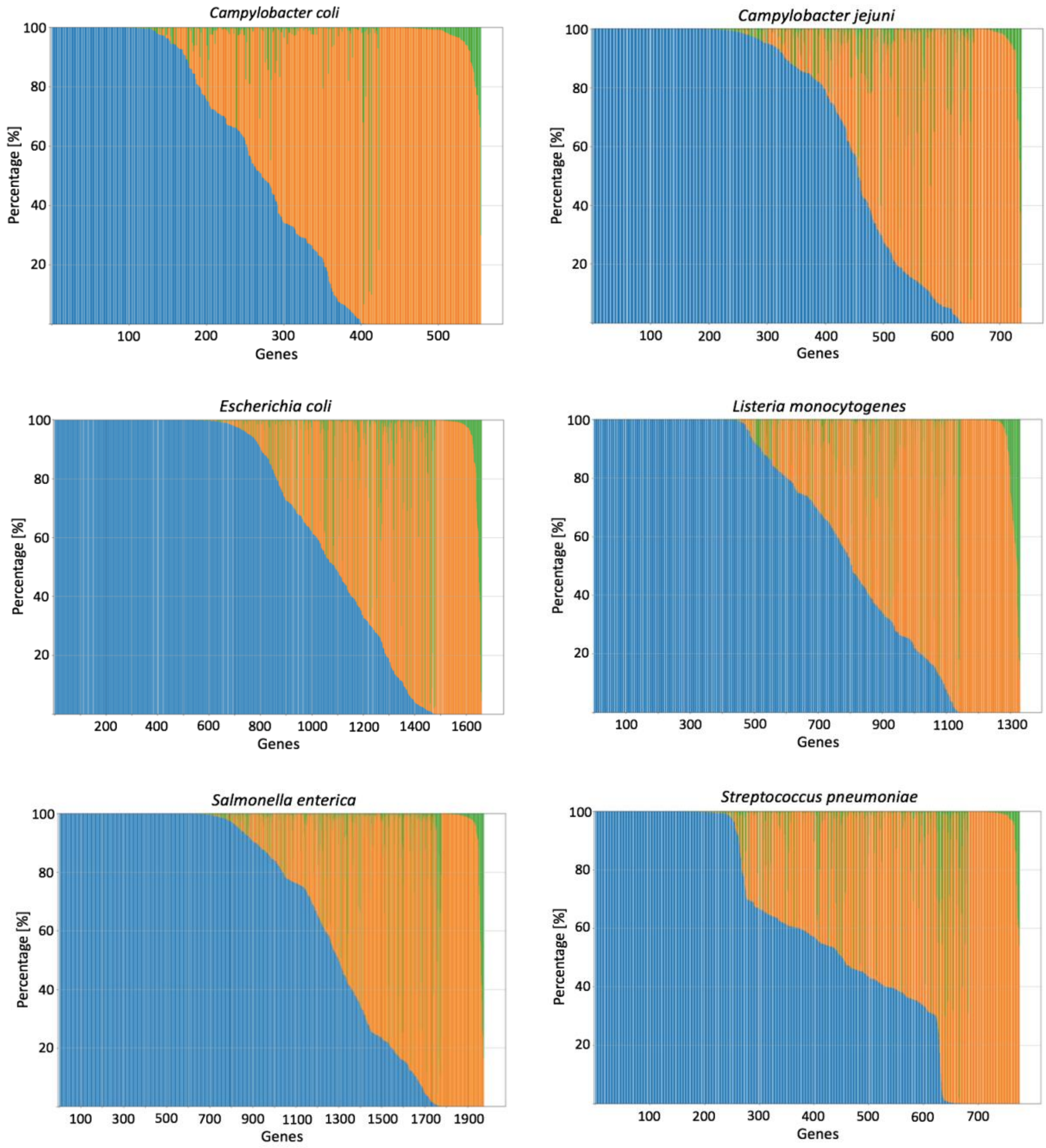


Figure 7. Visualisation of the conserved core gene pairs. The percentage is a measure for how often the upstream core gene in the core gene pair is directly followed by another core gene (blue), by accessory genes (orange), or by not classified genes (green). E.g. if a core gene has 50 % blue and 50 % orange, it means that the first core gene in the core gene pair was followed directly by the downstream core gene in 50 % of the genomes analysed, and directly followed by an accessory gene in the other 50 % of the genomes that was analysed.

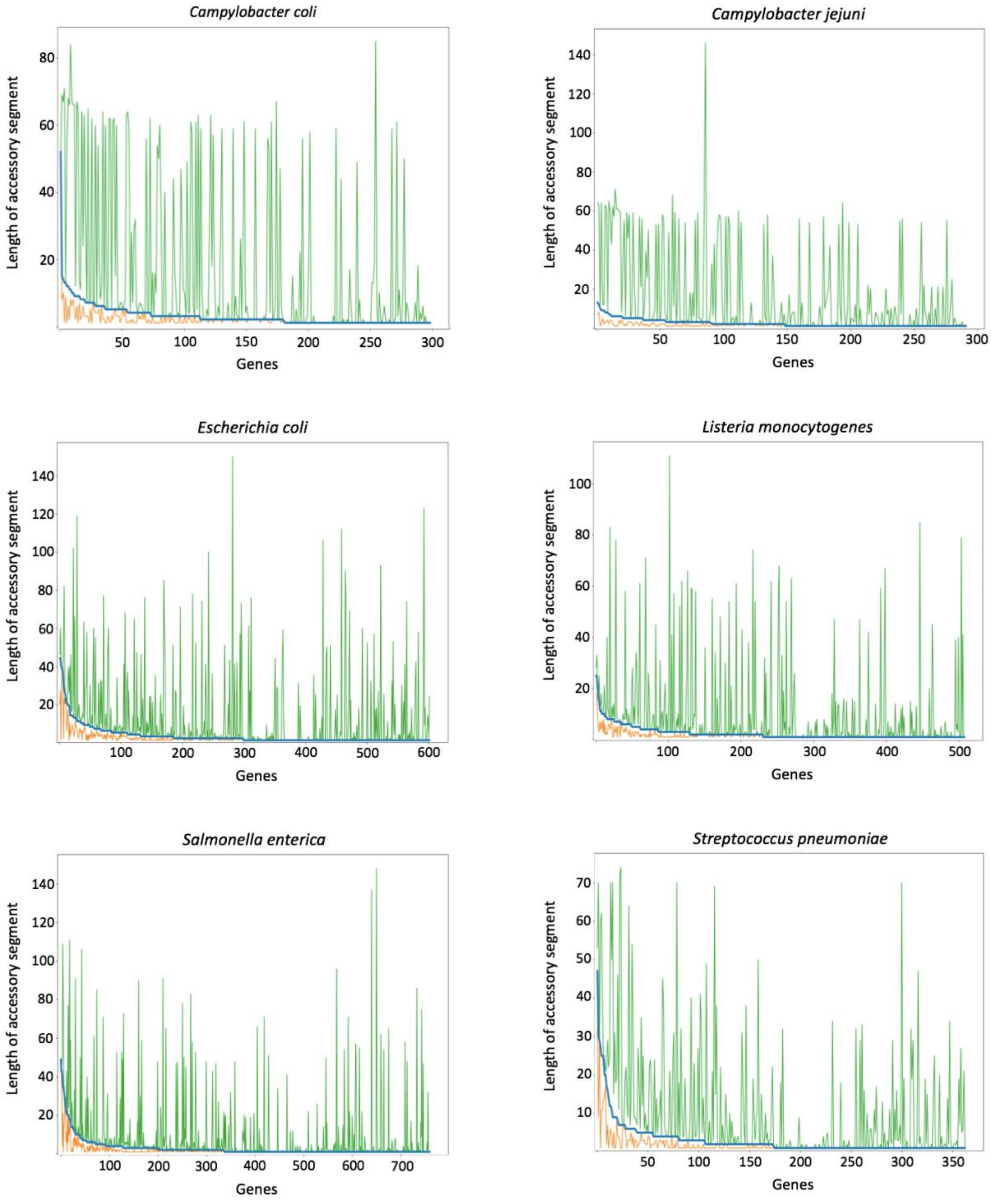


Figure 8. Length (number of genes) of the accessory segments inserted between two conserved core genes. For each gene, the median length (blue), maximum length (green), and minimum length (orange) for the accessory segments are shown.

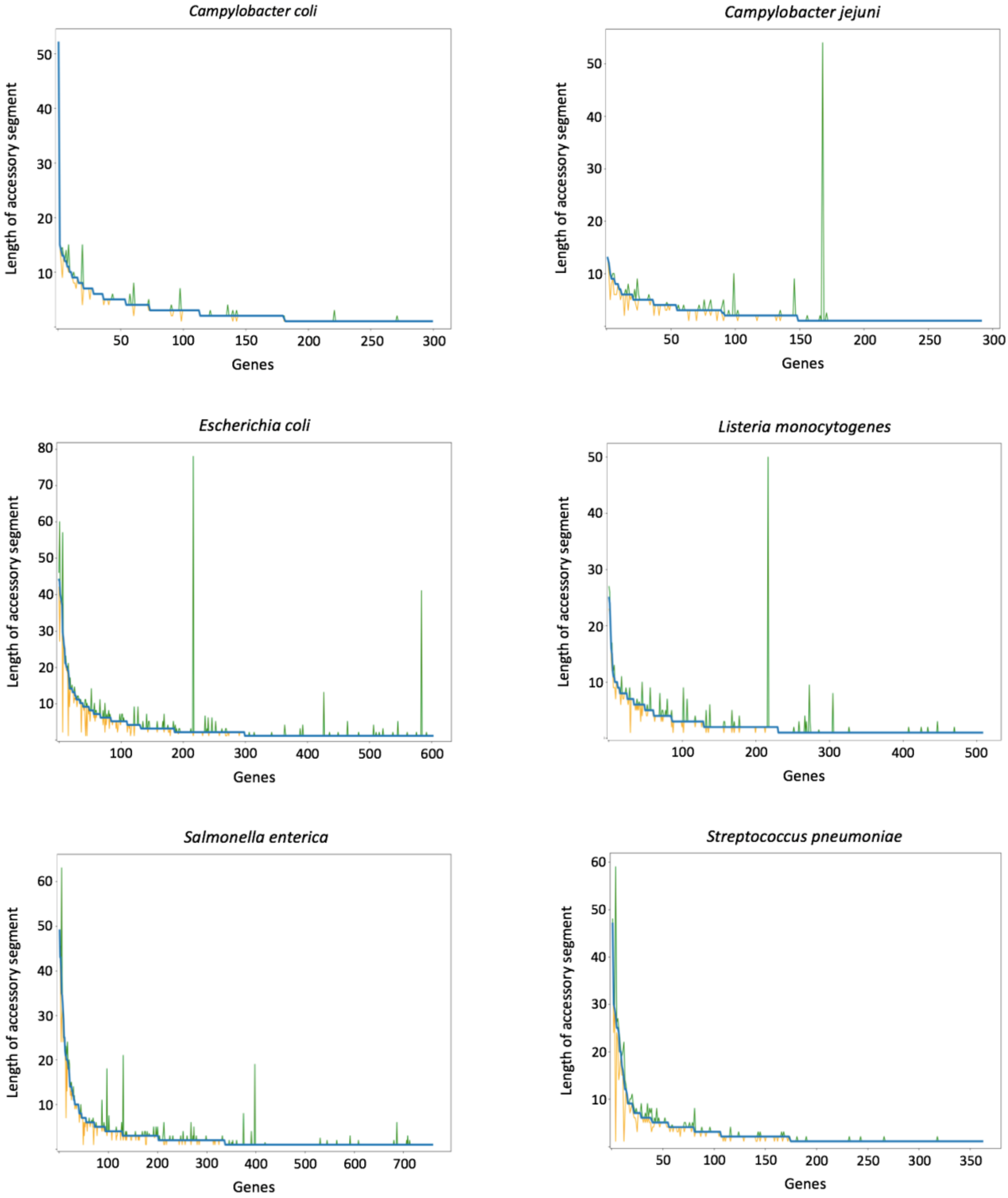


Figure 9. Length (number of genes) of the accessory segments inserted between two conserved core genes. For each gene, the median length (blue), third quartile (green), and first quartile (orange) for the accessory segments are shown.

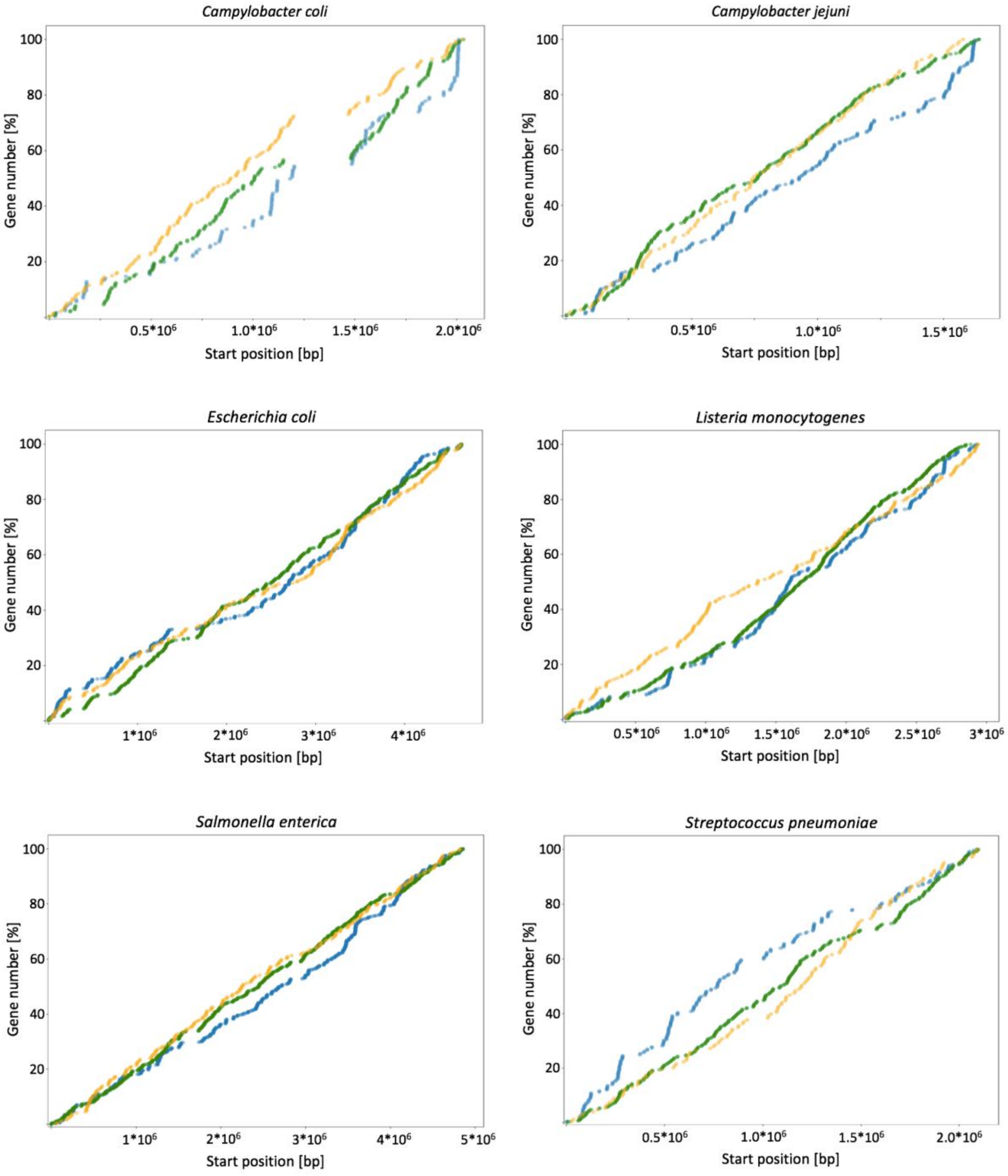


Figure 10. Cumulative plot of the distribution of the core genes' start position in their respective reference genomes. Each gene number represented as a percentage, making the percentage span the number of genes. Core genes directly followed by a core gene in at least 95% of the cases (blue), core genes directly followed by a core gene in 5%-95% of the cases (green), core genes directly followed by a core gene in less than 5% of the cases (orange).

5 Discussion

For many years, the number of genomes that are sequenced and available for analysis has grown exponentially. Especially for species included in surveillance projects, extensive numbers of genomes are accumulating. A few years ago, pan-genome analyses had only been made on small datasets using a small number of genomes. However, with the increasing number of genomes sequenced, the interest in making large pan-genome analyses is increasing too. Originally, publications with pan-genome studies were based only on a small number of genomes, having up to ten genomes in the datasets (Medini *et al.* 2005). Today publications are starting to show that larger datasets of up to thousands of genomes are used in pan-genome analyses (Park *et al.* 2019).

I chose to work with the six most common species from surveillance projects: *C. coli*, *C. jejuni*, *E. coli*, *L. monocytogenes*, *S. enterica*, and *S. pneumoniae*. The chosen species all had over 10,000 genomes available in the database, which were a sufficient amount of data to try to maximize the number of genomes taking part in the pan-genome analyses. The literature indicated that BPGA was the most powerful pan-genome tool available today for large-scale pan-genome analysis and it had been proven successfully for 1,000 genomes. Since there are a lot of sequenced genomes from surveillance projects, one aim in this thesis was to perform the analysis for as many genomes as possible for each species. I managed to successfully run the pan-genome analysis with BPGA for 1,000 genomes for *E. coli* and *S. enterica*, and 2,000 genomes for the other species. After this, a limit was reached. The limit for the number of genomes to be analysed was not reached because the RAM was limiting or that the computer ran out of capacity. This suggests that there is a need for further development of pan-genome analysis tools that can handle a larger amount of data than the tools available today. *E. coli* and *S. enterica* were the two species with the largest genome size, making it clear that the limit had to do with the genome size, and the number of genes to be analysed. It could be that the clustering step in the analysis only can handle a certain number of genes.

E. coli and *S. enterica* had more accessory-, and unique genes relative to their genome size compared to the other species. The shape of the cumulative growth curve for the pan-genome of *E. coli* visually indicated a higher openness of the genome than other species. At first, the cumulative growth curve for *E. coli* flattened out and almost reached a plateau, and then new genes were added at a higher rate again. This could be because the dataset consisted of different sub-populations. If this was the case, the first genomes that were entered into the analysis had a more similar gene content because they belonged to one sub-population, then other genomes were added later in the analysis which was not part of the sub-population. The openness of the pan-genomes was also estimated using Heaps' law, which supported that *E. coli* and *S. enterica* had more open pan-genomes than the other species. This could mean that *E. coli* and *S. enterica* has a higher tendency to pick up new genes from the environment and that they might have a more dynamic genome than the other species analysed in this thesis.

C. jejuni, *L. monocytogenes*, and *S. enterica* had the highest relative core genome size, indicating that these three species are the ones having the most conserved genomes among the analysed species. The lowest relative core genome size was found for *E. coli*, suggesting it to be the species with the least conserved genome. *C. coli* had a low relative core genome size too. However, it appeared through the analysis that most probably many core genes for *C. coli* were miss-classified as accessory genes due to a data artefact. If this would not have happened, *C. coli* would likely have a higher relative core genome size, close to the size of *C. jejuni*. Approximately half of the genes classified by BPGA were found in only one or very few genomes for all species. Species with a lot of unique genes or accessory genes found in very few genomes might be more prone to take up new genes. These low frequent genes might increase in the population to the next generation, especially if the genes are beneficial for the species in the environment it lives in. Or they might be phased out from the species if they instead are harmful.

An important aspect is that BPGA used a strict definition for core genes, where only the genes found in all genomes in the dataset analysed were classified as core genes. If a less strict definition for core genes were used during a pan-genome analysis instead, the number of core genes and the relative core genome size would increase for all species. The gene frequency distribution showed for all species that many accessory genes were found in more than 95 % of the genomes. These genes would instead be classified as core genes if a softer definition were used in a pan-genome analysis with a limit at 95 %. A strict definition can be sufficient when analysing small datasets and running pan-genome analyses for up to e.g. 20 genomes, which has originally been the case. However, when working with larger datasets having up to thousands of genomes, it could be more sufficient to use a softer limit for the core genes. This because when working with a larger number of genomes, there will be genes not found because of sequencing artefacts or because they were missed in the annotation step. In this thesis, I used a softer limit at 95 % as an example, however, this limit might need to be adjusted. Different limits might be suitable for different datasets depending on how the dataset was created. The limit may need adjustment depending on which sequencing method was used to generate the data, because the sequencing methods have different error rates. Another aspect to take into consideration can be which annotation method was used.

In the cumulative growth curves for the pan-genomes, it was seen that a lot of new genes were added for the first genomes in the pan-genome analyses. As more genomes were added, the curve flattened out as fewer new genes were found. Since the curves had flattened out considerably for all species, it means that most of the genes in the total dataset probably already were found in the analysis of the 1,000 or 2,000 genomes. This indicates that the number of genomes analysed was sufficient to include most of the genes from the species pan-genomes. However, if all genomes in the datasets were analysed, more unique genes and accessory genes present in only a few genomes would probably be found. *E. coli* were the species where the pan-genome curve had flattened out least, and it might have flattened out more if more genomes were analysed. One way to expand the analysis to a higher number of genomes even though BPGA could not handle a larger dataset would be to run BPGA several times for each species. For each run, a dataset of 1,000 genomes could be randomly chosen from the total dataset. After several runs, the results could be combined and thereby a more comprehensive analysis could be made. All species were estimated to have open pan-genomes during the pan-genome analysis by Heaps' law. For *C. jejuni*, *E. coli*, *S. enterica*, and *S. pneumoniae*, this confirmed what had been shown in previous studies (Halachev *et al.* 2011, Park *et al.* 2019). However, *L. monocytogenes* had been found to have a closed pan-genome previously (Halachev *et al.* 2011), which it was not during this thesis. Another round of pan-genome analysis with BPGA could have been made for *L. monocytogenes*, using 2,000 other genomes from its total dataset to see if the result would change. Although, it should be emphasized that the previously published article by Halachev *et al.* (2011) where *L. monocytogenes* was estimated to have a closed pan-genome, used the power law model to estimate the openness. Furthermore, they only had seven genomes in their analysis for *L. monocytogenes*. Because of their small number of genes in the pan-genome analysis, it is more likely that the pan-genome for *L. monocytogenes* is open. The power law model may be better to use than the Heaps' law to estimate the openness of pan-genomes. However, when using BPGA, the user could not choose which model to use.

For a functional classification of the genes, the genes were mapped towards COG categories and KEGG pathways. From the COG analysis, poorly characterized genes were the genes that could not get any functional classification among the genes present in the COG database. There was a larger portion of accessory-, and unique genes than core genes classified as poorly characterized. This shows that accessory-, and unique genes probably had more specific functions, which are less central. These functions have not been studied and thereby have not been added to COG or KEGG pathways. Core genes had a high percentage of their genes connected to metabolism in both COG categories and KEGG pathways. This was not very surprising since metabolism is connected to a lot of functions which is essential for all cells to survive. In KEGG pathways, a high percentage of the core genes were connected to genetic information processing. This is understandable since this category also includes a lot of necessary functions for cells survival, such as transcription, translation, replication, and reparation. Furthermore, accessory-, and unique genes had a higher percentage of its genes than the core genes classified as environmental information processing and as human diseases. Environmental information processing includes processes like membrane transport, signal transduction, signalling molecules, and interaction, which are aspects that could be affected by accessory-, and unique genes. This could have to do with specialised functions helping the bacteria handling the interaction with specific environments. The human diseases pathways include genes connected to e.g. drug resistance. Drug resistance genes can be picked up as unique genes, and when the bacteria live in an environment where it is advantageous to have this gene, it will spread through the population. The bacteria having the drug resistance gene might have a positive selection depending on the environment it lives in.

All species except *S. pneumoniae* had core genes showing a very high conservation-grade regarding their synteny. Some core genes had a conservation percentage at 100 %, and some almost 100 %. This means that some core genes had a very high, and some a total conservation regarding their synteny in the datasets analysed during this project. Furthermore, it was seen that a considerable number of core gene pairs were always placed directly after each other, not letting in any non-core genes between them. One reason why some core genes have such high synteny conservation and some core genes always sits directly after each other in the genome could be because they sit in operons. If another gene was inserted into an operon, the operon might lose its function, and therefore the core genes stay conserved and do always sit directly after each other in the genomes. Some core genes were instead always followed by non-core genes, the core genes in the core gene pairs were never sitting directly after each other. This shows that some areas in the genomes are very dynamic and open for taking in new genes between the core genes. In these areas, the core genes are probably not dependent on each other and do not need to sit directly after each other in the genome to function correctly.

It was shown that all species do take in considerably large accessory segments between core genes through their genomes. *E. coli* and *S. enterica* were the species with largest accessory segments. This increases the evidence that *E. coli* and *S. enterica* might have more dynamic genomes than the other species because they tend to have larger accessory segments and thereby let in more new genes between their core genes. Generally, for all species, the sites in the genomes where very large accessory segments were found, could be more dynamic and more open for taking in accessory segments than other sites in the genomes.

A comparison with a reference genome was made for all species to see how the core genes were distributed along the reference genome. The core genes were divided into three categories representing how often a gene was directly followed by another core gene or by accessory genes. Only the core genes were compared to the reference genomes since those are the genes expected to be found in the reference genomes. Even though the accessory genes were not compared to the reference genomes, it is possible to indirectly get a hint about where accessory genes are more likely to be found. This because of the different categories of core genes, where some core genes were always followed by accessory genes in the analysed genomes. All species had quite evenly distributed genes through the reference genomes for all categories of core genes. *E. coli* stood out from the other species with the most evenly distributed genes, i.e. it did not have any regions where different categories of core genes or accessory genes tended to be placed more often. The other five species all had some regions for different core gene categories where the gene density was higher. This indicated regions in the reference genome where different core genes and accessory genes sat closer together. Both *C. coli* and *C. jejuni* had a high gene density at the end of their reference genomes for the core genes always followed by another core gene. This could be because replication in bacteria is bidirectional, and thereby the core genes at the end of the genome are sitting close to *oriC* and will be replicated at the beginning of a replication fork. Because of this, the genes around *oriC* will be prioritized in the replication step, and the core genes could be placed here because they have essential, central functions for the cells survival. *C. coli* had a very large gap in the reference genome where none of the core genes was placed. Most likely, this indicates that the region only had non-core genes and held a very large accessory segment. The gap could also be a result of the wrongly annotated gene that took part in the pan-genome analysis for *C. coli*. It might be that a large part of the genes that would have been annotated as core genes if the *C. fetus* was not part of the analysis should have been placed where the gap is. There were some smaller gaps in the reference genomes for other species as well, which could be regions where accessory segments are placed too. To increase the reliability of the placement of the genes in the genomes, and to be able to draw confident conclusions, the placement of genes should be investigated further. For each species, several different reference genomes could be used to see where the genes are placed. In this thesis, only one reference genome was used for each species, which makes it hard to make validated conclusions about the genes exact placements in the genome. However, some indications can still be seen.

It was discovered that one genome in the *C. coli* dataset was not a *C. coli* but a *C. fetus*. This gave a gap in the pan-genome curve where a lot of new genes were added for this specific genome. Because of the wrongly annotated genome, the pan-genome analysis was affected a lot. Many genes that should have been classified as core genes were now most probably classified as accessory genes instead. Of course, this affects the number of core-, and accessory genes and the relative core genome size a lot for *C. coli*. If the wrongly annotated genome were not a part of the analysis, the number of core genes and the relative core genome size would be higher. Furthermore, there were two small gaps for *C. jejuni* as well. These genomes have not been investigated any further. Because of this, there is a risk that also these are wrongly annotated genomes, however, they could also just be genomes that contributed with a lot of new genes even though they are *C. jejuni* genomes. One reason why a genome of the same species could contribute with a higher number of new genes could be because they have a new plasmid that the genomes previously added to the analysis did not have. To prevent having wrongly annotated genomes in an analysis like this, it would be of good practice to have a pre-check of the genomes that goes through the database and make sure that the genomes are correctly annotated. It might be even more important in a project like this where the data were not annotated in the project but instead collected from different institutions. Different annotation pipelines contribute to a risk that some genomes have a lower quality of their annotations, which is a reason to why it could be good to annotate all genomes to be analysed with the same pipeline. If there would have been more time in this project, the wrongly annotated genome would have been replaced in the dataset and the analysis run again.

6 Conclusion

In this thesis, a pan-genome analysis was run with BPGA for six bacterial species. The pan-genome analyses ran successfully for a dataset of 1,000 genomes for *E. coli* and *S. enterica*, and 2,000 genomes for *C. coli*, *C. jejuni*, *L. monocytogenes*, and *S. pneumoniae*. All species were estimated to have open pan-genomes. From the functional annotation, it was seen that a large proportion of the core genes took part in central functions that are essential for the cells survival. The accessory genes, on the other hand, had a large proportion of its genes taking part in more specific functions which have not yet been studied. *C. jejuni*, *L. monocytogenes*, and *S. enterica* had the most conserved genomes among the studied species. *E. coli* and *S. enterica* had the most open and dynamic genomes, meaning they are more prone to take in new genes than the other species. Evidence showed that a portion of the core genes are placed in operons, and the operons are less dynamic and not open for letting in new genes. Other spots in the genomes were more dynamic and very prone to take in new genes. Most core genes were evenly distributed in the reference genomes for each species, although there were regions where the gene density was higher for the different categories of core genes too. Gaps in the reference genomes indicated spots where large accessory genomes might be placed. However, an extended analysis using more reference genomes should be done to draw confident conclusions from this. Finally, BPGA used a strict definition of the core genes, and in this project, it was shown that a softer limit for core genes might be more suitable for pan-genome analyses using large datasets.

To conclude, the pan-genome analysis could be successfully run for a large number of genomes for all species analysed during this project. It was shown that some species have more dynamic and more open genomes than others. Furthermore, it was shown that the dynamics in the genomes differ between different regions. Some regions in the genomes tend to more often take in accessory genes than others, indicating potential hotspots in the genomes.

7 Future Outlook

The work of this project could be expanded in several ways to get more knowledge about the core-, accessory-, and the pan-genome for the species analysed. Some of them are listed here:

- Use a pre-check to look through the genomes in the database used. Sort out the genomes that are wrongly annotated, and re-annotate them as the right species instead.
- Analyse larger datasets, either by running BPGA several times and then combining the results or by trying out another pan-genome tool. There is also the possibility to create a new software for analysing large datasets.

- Look more into specific genes which could be of interest. One example could be to look at the core gene pairs having the highest conservation percentage for the gene synteny.
- Make comparisons using several reference genomes to make more confident conclusions about the genes exact placement in the genomes.
- The gene-databases that were created for all species (explained in section 3.5) could be used for several new analyses that were not done in this project. One suggestion is to look at the annotations for all the classified genes and see if there are more e.g. drug-resistant genes found in the core genes or the accessory genes.
- More species could be added for the analysis to get more insight into other species monitored in surveillance projects.

8 Acknowledgements

First I would like to thank my supervisor Bo Segerman for helping me and giving me valuable insights through this project. I also want to thank my subject reader Dan I. Andersson for reviewing my report and work, and helping me to interpret some of my results biologically.

A special thanks to Lena Henriksson who coordinated this course, you are always so helpful, cheering, and solve all kinds of problems for us students. I appreciate this very much.

Lastly, thank you to all my family and friends who always support me and roots for me.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic Local Alignment Search Tool. Journal of Molecular Biology 215: 403–410.

Chaudhari NM, Gupta VK, Dutta C. 2016. BPGA- an ultra-fast pan-genome analysis pipeline. Scientific Reports, doi 10.1038/srep24373.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25: 1422–1423.

COG. 2021. Database of Clusters of Orthologous Genes (COGs). online January 2021: https://www.ncbi.nlm.nih.gov/research/cog. Accessed May 11, 2021.

Deng X, den Bakker HC, Hendriksen RS. 2016. Genomic Epidemiology: Whole-Genome-Sequencing–Powered Surveillance and Outbreak Investigation of Foodborne Bacterial Pathogens. Annual Review of Food Science and Technology 7: 353–374.

Halachev MR, Loman NJ, Pallen MJ. 2011. Calculating Orthologs in Bacteria and Archaea: A Divide and Conquer Approach. PLoS ONE 6: e28388.

Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. 2021. KEGG: integrating viruses and cellular organisms. Nucleic Acids Research 49: D545–D551.

Kanehisa M, Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. 4.

KEGG. 2021. KEGG PATHWAY Database. online May 1, 2021: https://www.genome.jp/kegg/pathway.html. Accessed May 11, 2021.

Lal D, May P, EuroEPINOMICS-RES Consortium, Perez-Palma E, Samocha KE, Kosmicki JA, Robinson EB, Møller RS, Krause R, Nürnberg P, Weckhuysen S, De Jonghe P, Guerrini R, Niestroj LM, Du J, Marini C, Ware JS, Kurki M, Gormley P, Tang S, Wu S, Biskup S, Poduri A, Neubauer BA, Koeleman BPC, Helbig KL, Weber YG, Helbig I, Majithia AR, Palotie A, Daly MJ. 2020. Gene family information facilitates variant interpretation and identification of disease-associated genes in neurodevelopmental disorders. Genome Medicine 12: 28.

Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. 2005. The microbial pan-genome. Current Opinion in Genetics & Development 15: 589–594.

Neville BA, O'Toole PW. 2014. MOLECULAR BIOLOGY | Genomics. Encyclopedia of Food Microbiology, pp. 770–779. Elsevier, Cork, Ireland.

Ozer EA, Allen JP, Hauser AR. 2014. Characterization of the core and accessory genomes of Pseudomonas aeruginosa using bioinformatic tools Spine and AGEnt. BMC Genomics 15: 737.

Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31: 3691–3693.

Park S-C, Lee K, Kim YO, Won S, Chun J. 2019. Large-Scale Genomics Reveals the Genetic Characteristics of Seven Species and Importance of Phylogenetic Distance for Estimating Pan-Genome Size. Frontiers in Microbiology 10: 834.

Tatusov RL. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Research 28: 33–36.

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, DeBoy RT, Davidsen TM, Mora M, Scarselli M, Immaculada Margarit y Ros, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, Kevin J. B. O'Connor, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM. 2005. Genome Analysis of Multiple Pathogenic Isolates of Streptococcus agalactiae: Implications for the Microbial "Pan-Genome." Proceedings of the National Academy of Sciences - PNAS 102: 13950–13955.

Tettelin H, Riley D, Cattuto C, Medini D. 2008. Comparative genomics: the bacterial pan-genome. Current Opinion in Microbiology 11: 472–477.

Van Rossum G, Drake FL. 2009. Python 3 Reference Manual. CreateSpace, Scotts Valley, CA.

Vernikos GS. 2020. A Review of Pangenome Tools and Recent Studies. In: Tettelin H, Medini D (ed.). The Pangenome: Diversity, Dynamics and Evolution of Genomes, pp. 89–112. Springer International Publishing, Cham.

WHO. 2021. Foodborne diseases. online 2021: https://www.who.int/health-topics/foodborne-diseases#tab=tab_1. Accessed May 11, 2021.

Appendix A – KEGG pathway categories

The category *metabolism* consists of the following under categories:

- Global and overview maps
- Carbohydrate metabolism
- Energy metabolism
- Lipid metabolism
- Nucleotide metabolism
- Amino acid metabolism
- Metabolism of other amino acids
- Glycan biosynthesis and metabolism
- Metabolism of cofactors and vitamins
- Metabolism of terpenoids and polyketides
- Biosynthesis of other secondary metabolites
- Xenobiotics biodegradation and metabolism
- Chemical structure transformation maps

The category *genetic information processing* consists of the following under categories:

- Transcription
- Translation
- Folding, sorting and degradation
- Replication and repair

The category *environmental information processing* consists of the following under categories:

- Membrane transport
- Signal transduction
- Signalling molecules and interaction

The category *cellular processes* consists of the following under categories:

- Transport and catabolism
- Cell growth and death
- Cellular community – eukaryotes
- Cellular community – prokaryotes
- Cell motility

The category *organismal systems* consists of the following under categories:

- Immune system
- Endocrine system
- Circulatory system
- Digestive system
- Excretory system
- Nervous system
- Sensory system
- Development and regeneration
- Aging
- Environmental adaptation

The category *human diseases* consists of the following under categories:

- Cancer: overview
- Cancer: specific types
- Infectious disease: viral
- Infectious disease: bacterial
- Infectious disease: parasitic
- Immune disease
- Neurogenerative disease
- Substance dependence
- Cardiovascular disease
- Endocrine and metabolic disease
- Drug resistance: antimicrobial
- Drug resistance: antineoplastic

Appendix B – COG distributions

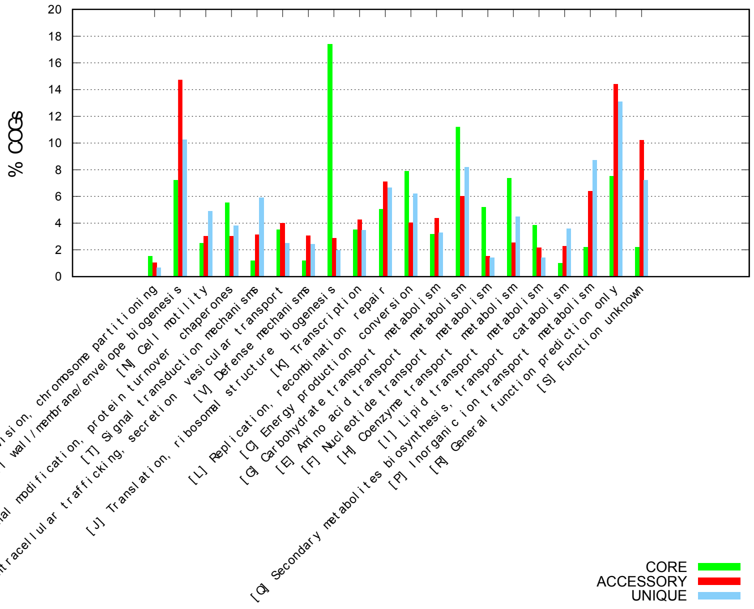


Figure A1. COG distribution of genes classified by BPGA for *Campylobacter coli.*

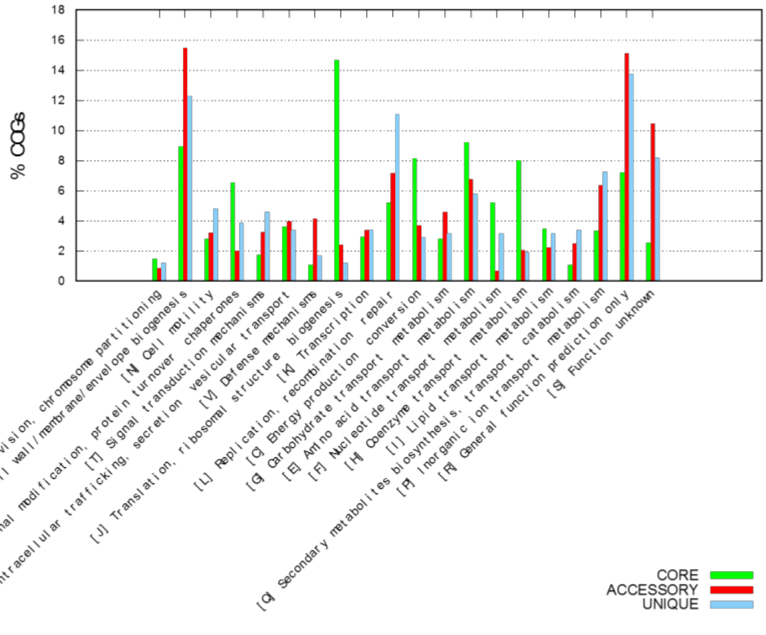


Figure A1. COG distribution of genes classified by BPGA for *Campylobacter jejuni.*

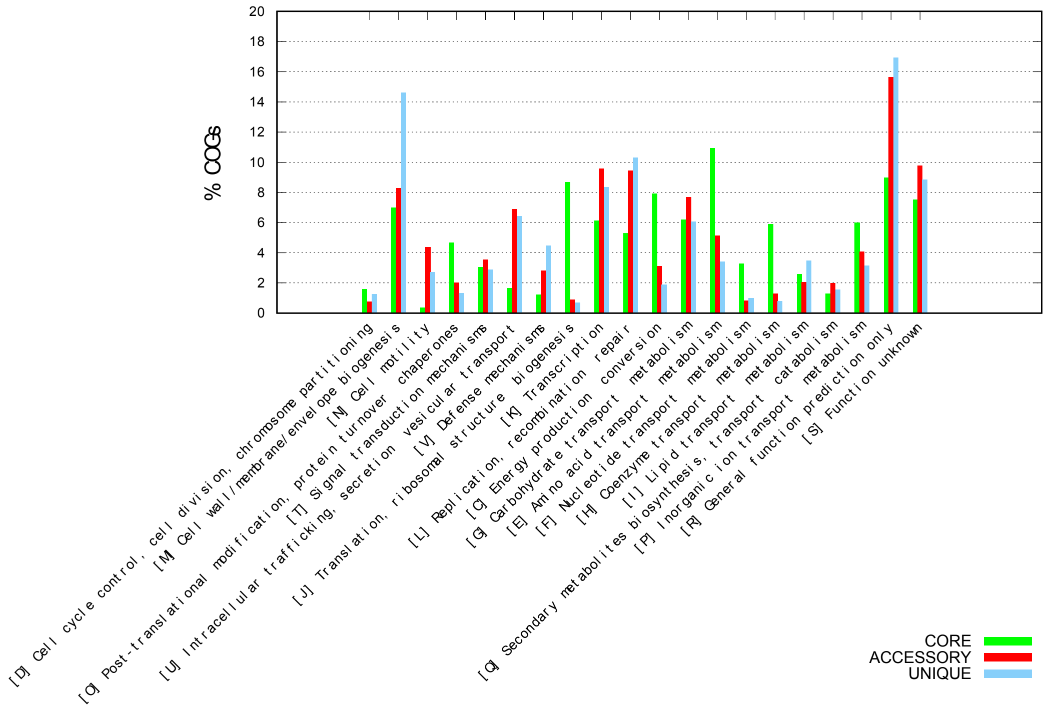


Figure A2. COG distribution of genes classified by BPGA for *Escherichia coli.*

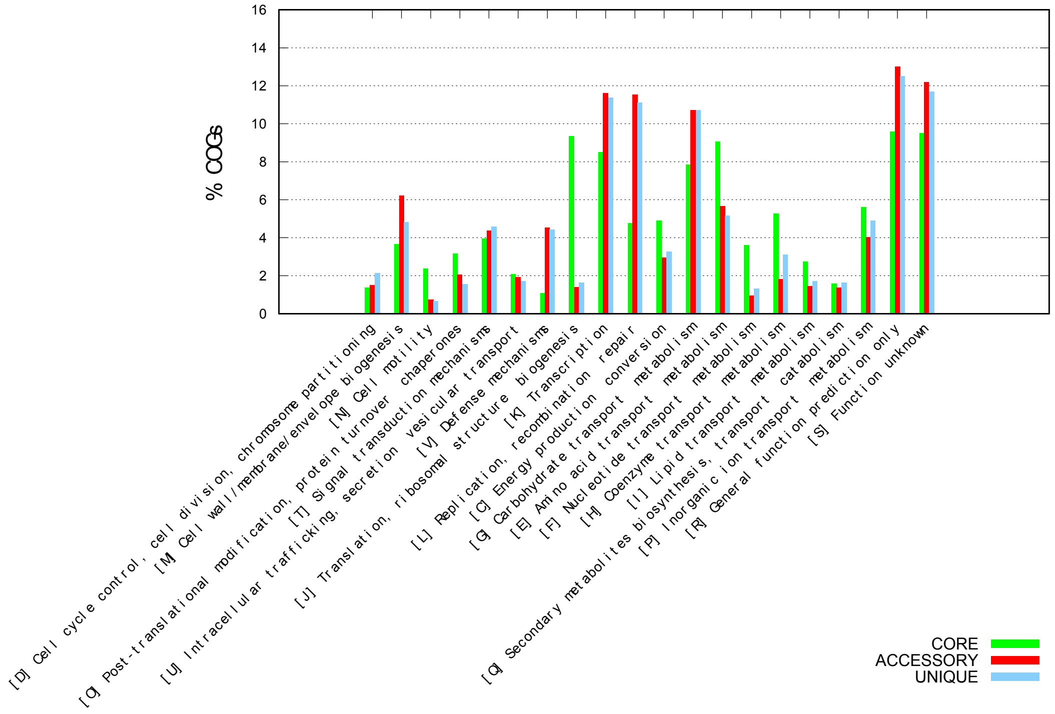


Figure A3. COG distribution of genes classified by BPGA for *Listeria monocytogenes.*

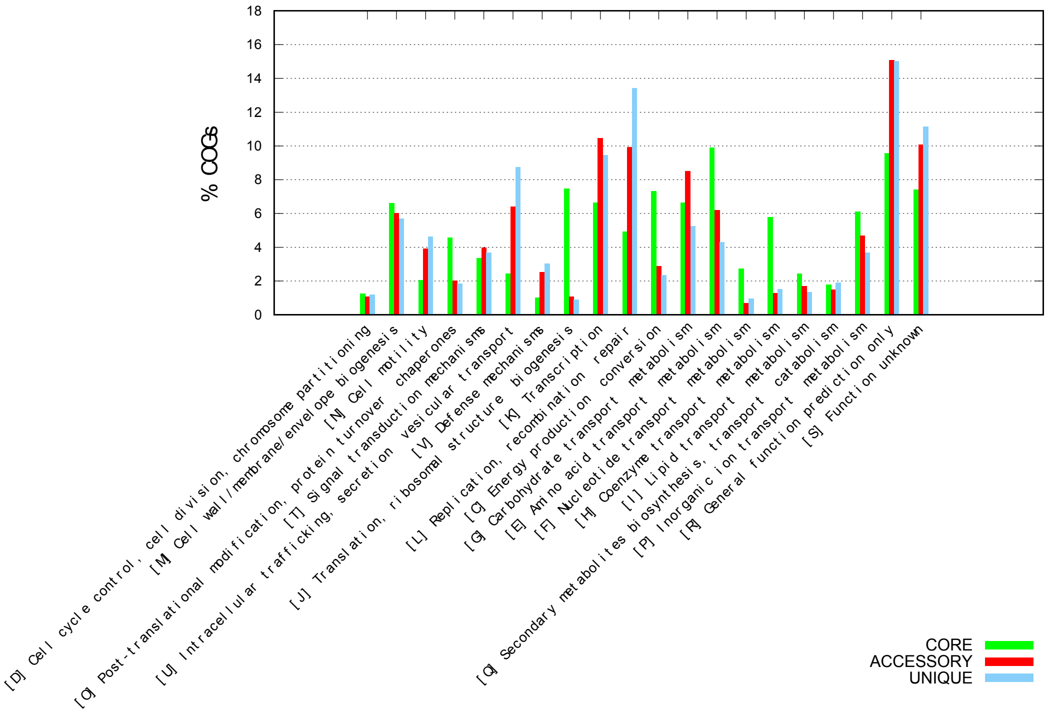


Figure A4. COG distribution of genes classified by BPGA for *Salmonella enterica.*

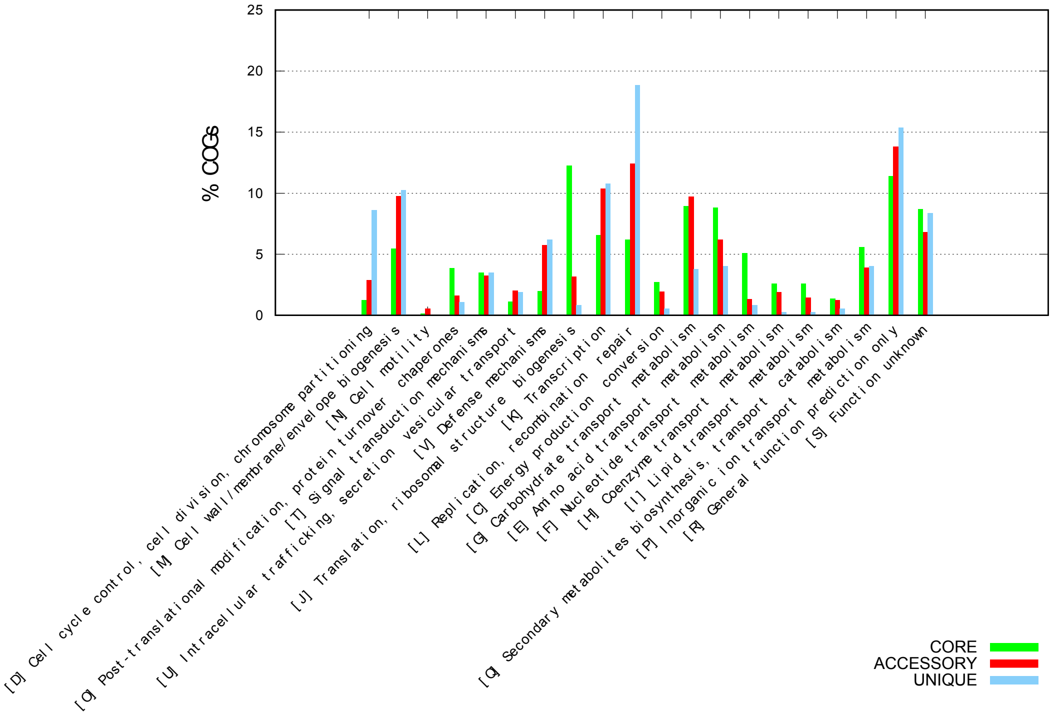


Figure A5. COG distribution of genes classified by BPGA for *Streptococcus pneumoniae.*

Appendix C – KEGG distributions

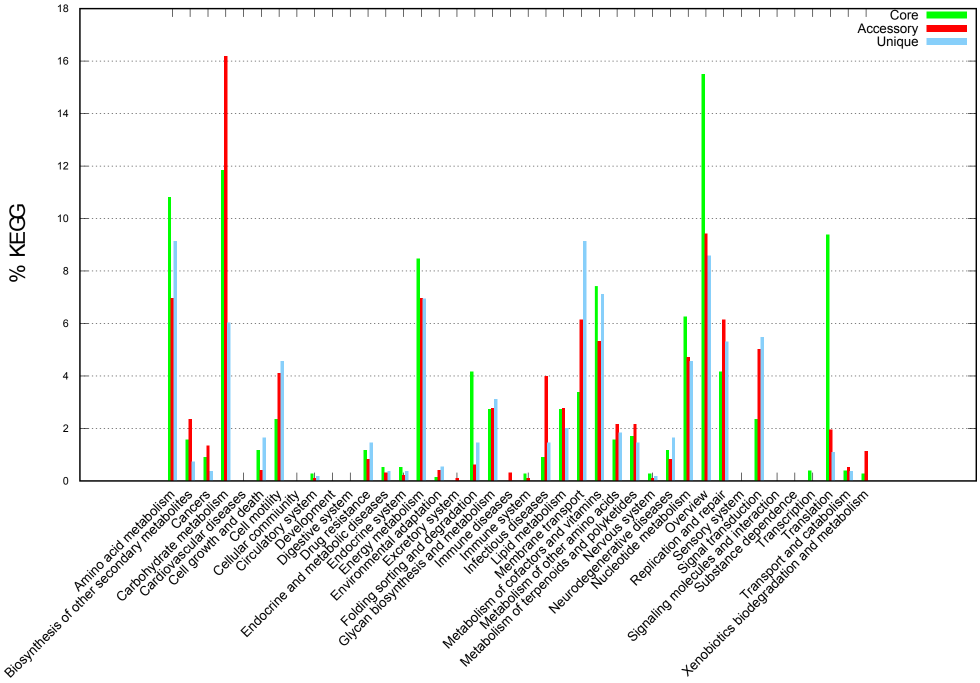


Figure A6. KEGG distribution of genes classified by BPGA for *Campylobacter coli.*

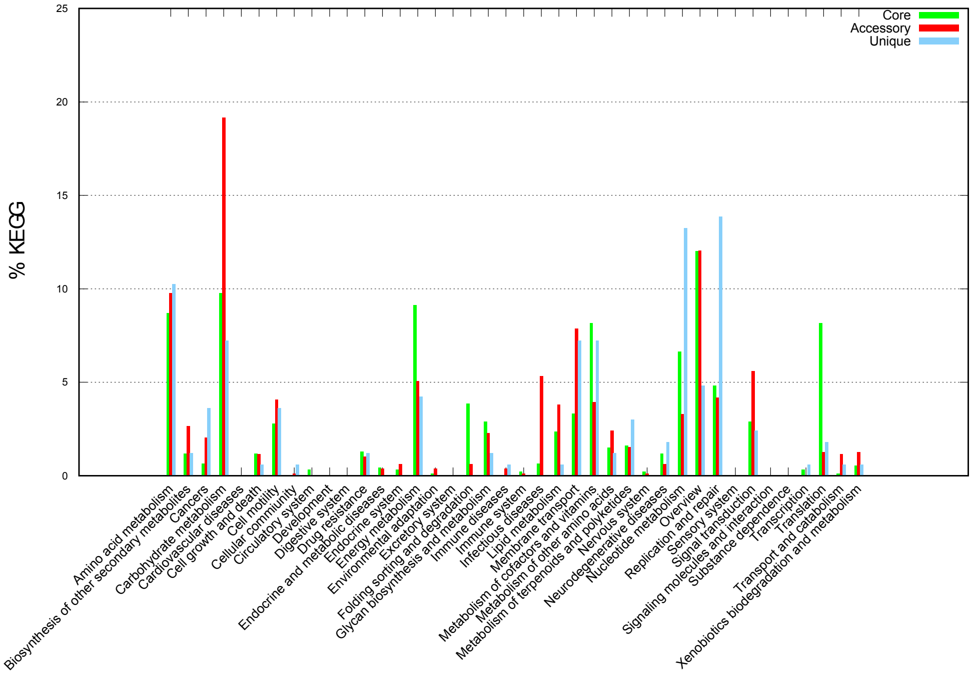


Figure A7. KEGG distribution of genes classified by BPGA for *Campylobacter jejuni.*

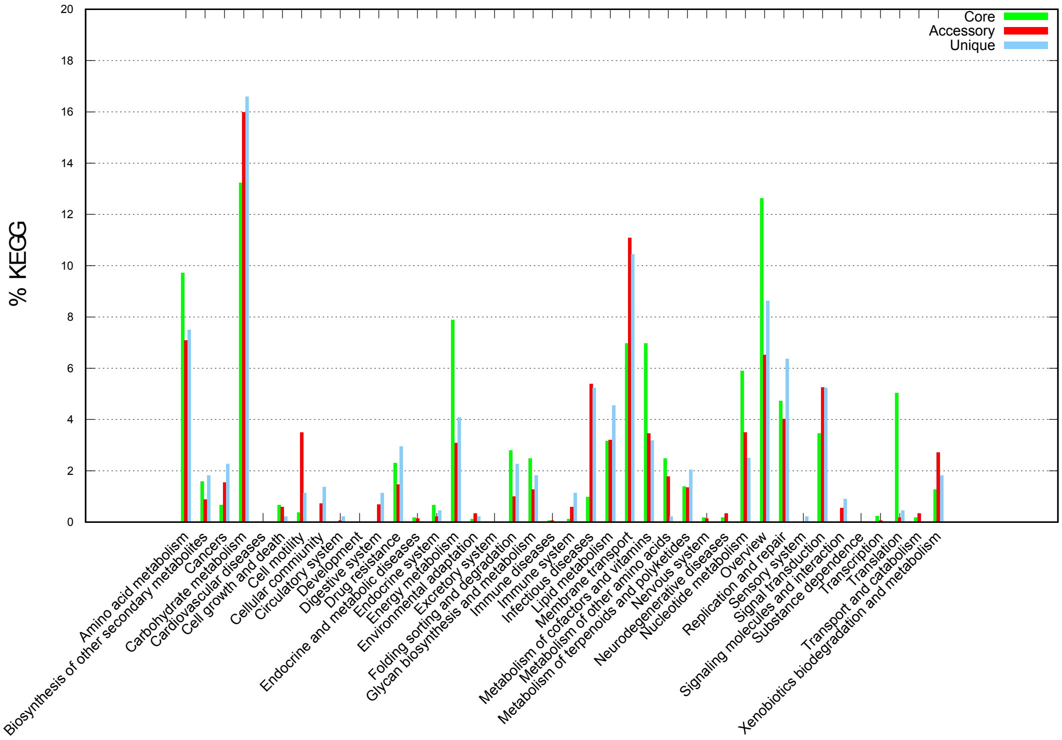


Figure A8. KEGG distribution of genes classified by BPGA for *Escherichia coli.*

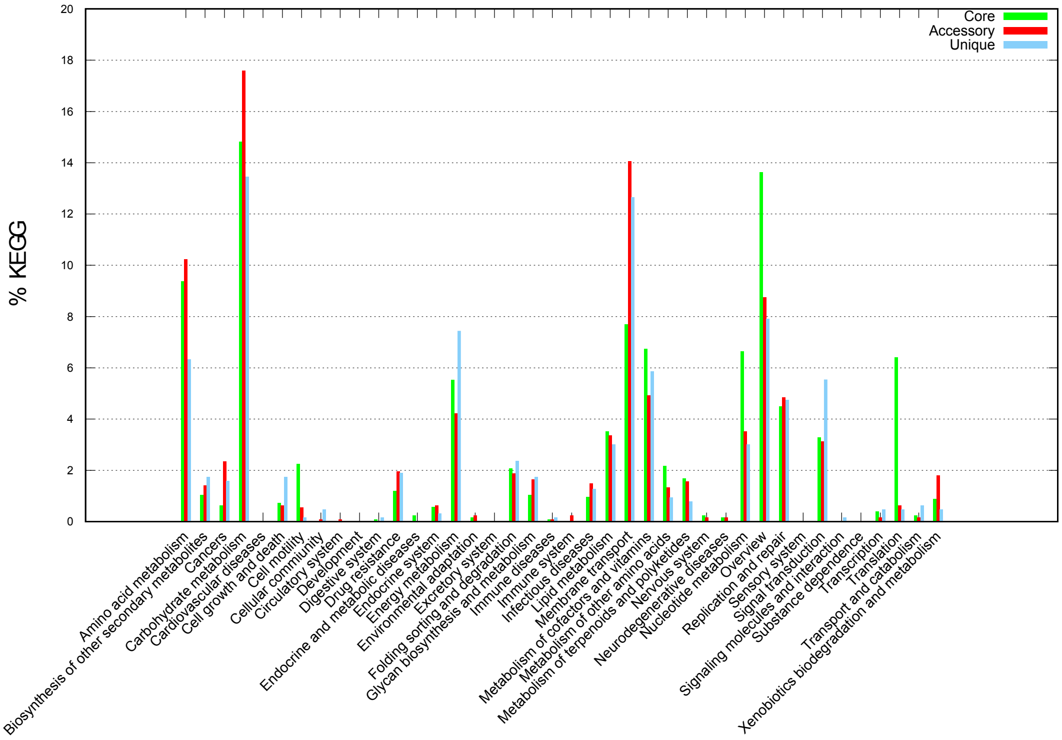


Figure A9. KEGG distribution of genes classified by BPGA for *Listeria monocytogenes***.**

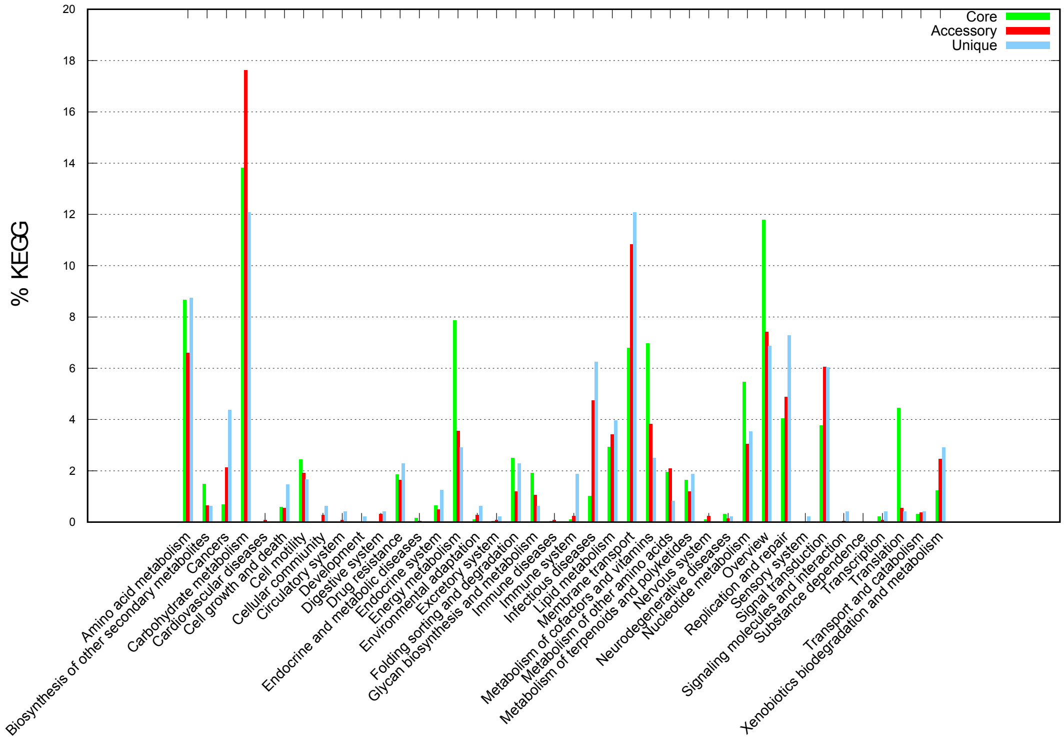


Figure A10. KEGG distribution of genes classified by BPGA for *Salmonella enterica.*

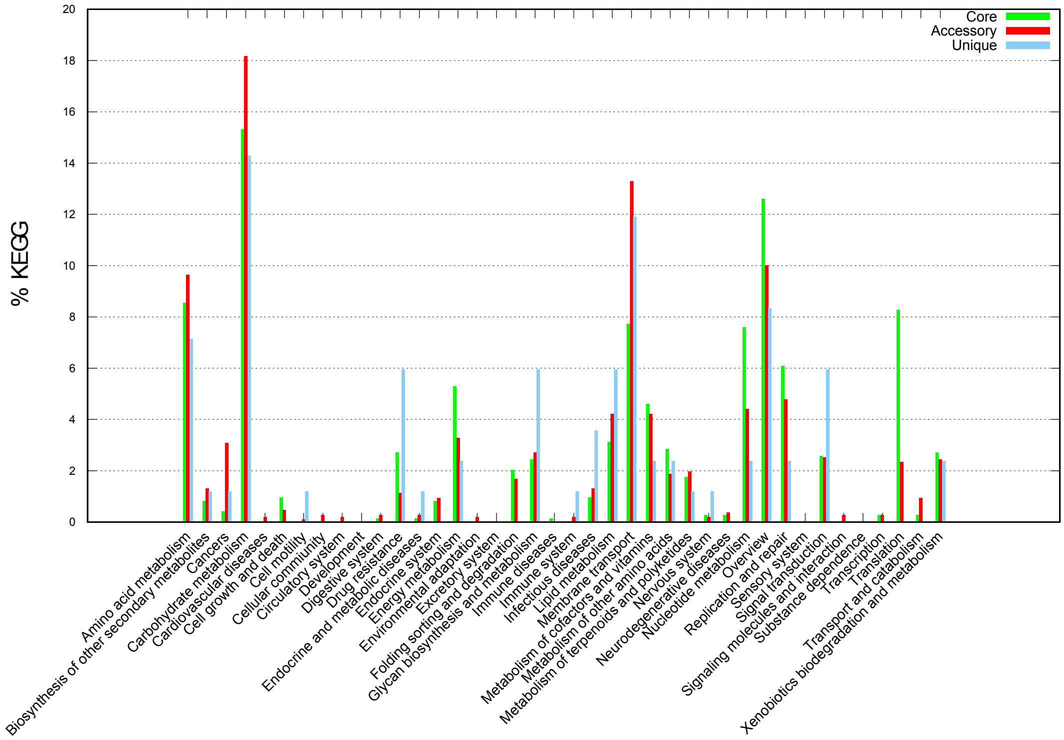


Figure A11. KEGG distribution of genes classified by BPGA for *Streptococcus pneumoniae.*