Methods in Molecular Biology

Comparative Genomics, Volume 2

Chapter: Step-by-step Bacterial Genome Comparison

Authors: Dennis Carhuaricra-Huaman and João Carlos Setubal

Summary/Abstract

Thousands of bacterial genome sequences are available in public databases thanks to advances in genome sequencing and bioinformatics in the last decade. Comparative genomics methods have allowed explore bacterial diversity and understand their evolution. In this method protocol, we describe a complete bioinformatic workflow for comparative genomics including genome annotation, pangenome

reconstruction and visualization, phylogenetic analysis and identification of sequences of interest such as antimicrobial-resistant genes, virulence factors and phage sequences using state-of-the-art, open-source tools. Furthermore, we present a case study of comparative analysis of *Salmonella enterica* serovar Typhimurium to illustrate our protocol providing Linux command lines and scripts to generate high-quality visualization in R environment. The presented workflow provides a user-friendly protocol that researchers with basic expertise in bioinformatics can easily follow to conduct comparable investigations.

Key Words: Comparative genomics, pangenome, prokaryotes, phylogenetic analysis

1. Introduction

Bacterial genomics started in 1995 with the publication of the *Haemophilus influenzae* complete genome enabling a more profound comprehension of the organism's biology [1]. Very shortly after, the *Mycoplasma genitalium* genome was sequenced. The availability of these two genomes gave rise to comparative genomics, revolutionizing prokaryotic biology [2]. Comparative genomics enables many analyses, the most basic of which is the determination of which genes are present or absent in a particular genome with respect to others. Such information helps the understanding of genome evolution and the basis of phenotypic differences among related organisms.

Early comparative studies already showed that large differences in gene content may occur between genomes of the same prokaryotic species. When three *Escherichia coli* genomes became available in 2002, the comparative analysis revealed that only 39.2 % of the genes were shared by these three genomes [3]. Soon other similar observations were made about other bacterial species, showing the remarkable plasticity of such genomes, eventually giving rise to the pangenome concept. The

pangenome is the set of all non-redundant genes present in a given set of genomes. The differential gene content among genomes from the same species comes about because of extensive horizontal gene transfer (HGT) and gene loss, two of the main forces driving the evolution of prokaryotes [4].

In the last twenty years, the development of cheaper and more accessible sequencing methods caused an exponential increase in the amount of bacterial genomic data [5]. This in turn stimulated the development of genome informatics, or computational methods for genome analysis, in particular comparative analysis. A few examples are the genomic epidemiology of pathogenic bacteria [6], pathogenesis and niche specialization [7], discovery of genes associated with virulence and antimicrobial resistance [8, 9], identification of antigens through reverse vaccinology [10], discovery of antiphage defense systems [11], among others.

The use of massive sequencing and large genomic datasets may lead to important discoveries in various fields, such as medicine and plant pathology. However, frequently there is a mismatch between the capability of obtaining large genomic datasets and the ability of effectively analyzing such data. Genome sequences are only useful if there are adequate capabilities for annotation and comparative analysis, including computational infrastructure and skilled scientists in data analysis and programming.

In this chapter, we describe the step-by-step genomic comparisons using the bacteria *Salmonella enterica* serovar Typhimurium (*S.* Typhimurium) as an example. *S.* Typhimurium is one of the most important gastrointestinal pathogens of humans and is carried by livestock. In this chapter we use genomes of *S.* Typhimurium downloaded from a public database. The protocol presented in this study uses state-of-the-art bioinformatic tools that are freely available. This protocol begins with genome annotation and pangenome reconstruction, followed by gene content analysis and visualization, and phylogenetic reconstruction. The protocol concludes with the identification of sequences of interest, including antimicrobial-resistant genes (ARGs), virulence genes, and phage sequences. The aim of this

protocol is to provide a user-friendly guide that can be used as a template by researchers who are interested in applying the same analyses to their own genome datasets.

2. Requirements and Assumptions

This chapter assumes basic knowledge of Unix/Linux and R. All analyses can be run on a desktop computer running Linux/Unix or Mac OSX. Most programs can be executed using bash shell commands. We adopt the convention of presenting commands executed on the Linux shell preceded by the "$" symbol and the label **bash shell**. We also present R code, which can be executed in the RStudio environment (https://posit.co/download/rstudio-desktop/). R code sections are preceded by the label **R script**.

3. Datasets

Bacterial genome sequences can be retrieved from a variety of public repositories; examples are GenBank [12], BIGSdb [13], IMG [14], BV-BRC [15] or Enterobase [16]. In this chapter, we will use genome sequences of *Salmonella enterica* ser. Typhimurium, which we abbreviate as **SeT**. Our dataset consists of 12 genomes from strains LT2, 798, D23580, DT104, DT2, L-3553, SL1344, T000240, U288, SO4698-09, SO9207-07, SO9304-02, which have been isolated from different hosts [17], and two SeT genomes isolated from guinea pigs (SMVET11, SMVET22) sequenced by our group [18]. The accessions of these 14 SeT genomes are shown in Table 1 and can be downloaded from the Genbank genome database. Table 1 also contains additional information associated with each sample, such as host source, country, and genetic characteristics that were identified in the mentioned studies, such as the sequence type (ST), a method based on the allelic profile of seven housekeeping genes; and the phylogroup, which denotes a strain's placement in the SeT phylogenetic tree, in clades α or β [17].

Table 1. The genomes of *S.* Typhimurium used in this chapter. Each genome sequence can be downloaded from GenBank using the Assembly identifier in the last column.

Genome	Host	Country	MLST	Phylogroup	Assembly
798	Pig	USA	ST19	β	GCA_000252875.1
D23580	Human	Africa	ST313	β	GCA_000027025.1
DT104	Cattle	NA	ST19	α	GCA_000493675.1
DT2	Pigeon	Germany	ST128	β	GCA_000493535.2
L-3553	Cattle	Japan	ST19	β	GCA_000828595.1
LT2	Human	USA	ST19	α	GCA_000006945.2
SL1344	Cattle	UK	ST19	β	GCA_000210855.2
T000240	Human	Japan	ST19	α	GCA_000283735.1
U288	Pig	UK	ST19	α	GCA_000380325.1
SO4698-09	Cattle	UK	ST34	α	GCA_001540845.1
SO9207-07	Pig	UK	ST19	α	GCA_903989485.1
SO9304-02	Cattle	UK	ST19	β	GCA_902500315.1
SMVET11	Guinea Pig	Peru	ST19	α	GCA_024721515.1
SMVET22	Guinea Pig	Peru	ST19	α	GCA_024721395.1

4. **Software**

The software tools that will be used in this workflow are freely available and summarized in Table 2, including links to the websites from which these can be downloaded and installed.

Table 2: List of software tools used to perform comparative genomics in this chapter

Tool	**Description**	**Ref**	**Source**
abricate	Genome screening for ARGs and VFs	[19]	https://github.com/tseemann/abricate
Easyfig	Genome sequence comparison	[20]	https://mjsull.github.io/Easyfig/
egg-NOG mapper v.2	Funtional enrichment	[21]	https://github.com/eggnogdb/eggnog-mapper
Genbank	Genome database	[12]	https://www.ncbi.nlm.nih.gov/genbank/
Gephi	Network visualization	[22]	https://gephi.org/users/download/
ggplot2	R package	[23]	https://cran.r-project.org/web/packages/ggplot2/
ggtree	R package	[24]	https://github.com/YuLab-SMU/ggtree
Gubbins	Recombination	[25]	https://github.com/nickjcroucher/gubbins
IQ-TREE 2	Phylogeny	[26]	https://github.com/iqtree/iqtree2
limma	R package	[27]	https://bioconductor.org/packages/release/bioc/html/limma.html

micropan	R package	[28]	https://github.com/larssnip/micropan
Panaroo	Pangenomic analysis	[29]	https://github.com/gtonkinhill/panaroo
PPanGGolin	Pangenome graph	[30]	https://github.com/labgem/PPanGGOLiN
pheatmap	R package	[31]	https://github.com/cran/pheatmap
Prokka	Genome Annotation	[32]	https://github.com/tseemann/prokka
R	Environment for analysis	[33]	https://www.r-project.org/
Scoary	Pangenome-wide association studies analysis	[34]	https://github.com/AdmiralenOla/Scoary
vegan	R package	[35]	https://github.com/vegandevs/vegan
virsorter2	Phage sequence prediction	[36]	https://github.com/jiarong/VirSorter2

5. Genome Annotation

Genome annotation is the process whereby the location and functional characteristics of genes and other genetic elements are added to the raw genome sequence. Nowadays, it is an automated process, with occasional manual curation in special cases and/or for particular genes. With more than a million prokaryotic genomes available in public databases, accurate genome annotation is crucial for many downstream genomic analyses [37]. **Prokka** [32] is the most cited command-line tool for prokaryote genome annotation, with an easy installation and short runtime performance (5 minutes for a typical bacterial genome of about 5 Mbp). The prokka pipeline uses prodigal [38] to predict coding sequences (CDSs) and other software tools for RNA annotation. Recently, **Bakta** [39] was introduced using a similar workflow but providing a more comprehensive annotation. Bakta is capable of predicting pseudogenes

and small proteins that are not annotated by Prokka. Both tools provide a variety of output files, such as .gff, .gbk and .faa files, which are commonly used for comparative analysis.

Here, we annotate all 14 genomes of our dataset with prokka. We loop through all of the genomes using the following command (which assumes that the FASTA files for all genomes are available in the directory from which the command is issued):

Bash shell

```
$ for i in *.fasta; do prokka --kingdom Bacteria --genus
Salmonella --prefix "${i%%.*}" --locustag "${i%%.*}" --outdir
"${i%%.*}" --compliant "$i"; done
```

Here we are asking prokka to annotate all assemblies in FASTA format (.fasta); name the output files (--prefix) and locus tag (--locustag) as the isolate name; and make the annotations compliant with NCBI standards (--compliant).

For each genome annotated, a directory with the same name is created containing annotation files. Three output files (.gff, .gbk and .faa) will be used for downstream analysis. Table 3 summarizes the annotation features of all 14 genomes.

Table 3. Annotation statistics of SeT genomes. In the last column, the number represents the total number of rRNA gene units that were detected (among three possible: 5S, 23S, 16S).

Genome	Size (bp)	# CDS	# tRNA	# rRNA units
798	4970096	4685	83	22
D23580	4879400	4554	88	22
DT104	5027665	4743	85	22

DT2	4814801	4583	84	22
L-3553	5184452	4925	85	22
LT2	4951383	4620	86	22
SL1344	5067450	4763	86	22
T000240	5069994	4784	84	22
U288	5017059	4707	85	22
SO4698-09	5037238	4750	85	22
SO9207-07	4916754	4585	88	22
SO9304-02	5045986	4789	86	22
SMVET11	4851410	4542	76	8
SMVET22	5095938	4863	75	8

We now describe the tool eggNOG-mapper, which can map protein-coding genes, or coding sequences (CDSs), in a genome to the orthologous families in the eggNOG database [40]. eggNOG mapper offers functional annotation that includes KEGG pathways [41], COG functional categories [42], carbohydrate-active enzymes (CAZymes) families [43] and Gene Ontology terms [44]. Protein sequences (.faa file) generated by prokka can be uploaded to the eggNOG-mapper website (http://eggnog-mapper.embl.de/), which uses precomputed orthologous groups from the eggNOG database v.6 (http://eggnog6.embl.de) for fast functional annotation. eggNOG-mapper can also be installed locally, but the storage requirement is high due to database size (around 40 Gb is needed for the eggNOG annotation databases and additional disk space if the HMMER option is used). eggNOGG-mapper can be executed locally with the following command:

Bash shell

```
$ for i in *.faa; do emapper.py -i "$i" --output "$i"_eggnog;
done
```

For each input, eggNOGG-maper returns an annotation file (*samplename.emapper.annotations*) which provides the predictions for each query in TSV format (tab-separated values). Relevant columns are as follows:

Query: the query sequence name

GOs: list of predicted GO terms

COG_category: list of predicted COG categories

KEGG_ko: list of predicted KEGG orthologs

CAZy: list of predicted CAZy orthologs

6. Pangenome Reconstruction and Visualization

The pangenome is the set of all gene families present in a determined group of genomes belonging to a specific taxon [45]. The gene set can be subdivided into the "core" and "accessory" genomes. The core genome is composed of genes that are present in all members of the group, whereas those genes that are only present in some members represent the accessory genome [46]. Depending on the program used to compute the pangenome, the categories 'core' and 'accessory' can be further subdivided (as will be the case in the example of this chapter, as shown below). This type of analysis is restricted to protein-coding genes.

6.1 Ortholog gene computation and clustering

Identifying orthologous genes in different genomes is the first step for pangenome reconstruction and gene content comparison. Orthologs can be identified by using carrying out similarity searches between

genes from different genomes, using for example BLAST [47], CD-HIT [48], or DIAMOND [49], and then clustering the results into orthologous groups using the Markov Clustering algorithm (MCL) or by looking at triangles of pairwise best hits [50, 51].

Panaroo [29] is currently one of the most popular tools for pangenome reconstruction and is the tool we use in this example. Given a set of annotated genomes, in the form of .gff or .gbk files, panaroo uses CD-HIT for sequence similarity search and clustering in order to obtain gene clusters with a high similarity threshold (98% by default). Some of these clusters are then merged according to synteny information, which is also used to find missing genes and correct for possible errors in assembly and annotation.

In order to reconstruct the pangenome of our dataset of SeT genomes, the annotation files (.gff format) previously generated by prokka are used as input to panaroo. The basic command for pangenome calculation is as follows:

Bash shell

```
$ panaroo -i *.gff -o results_SeT --clean-mode strict -a core
--aligner mafft
```

The parameter 'aligner' indicates the program that panaroo should use to perform multiple alignment of the genes in each cluster. In this case we use the program MAFFT [52]. Panaroo will create a directory called 'results_ST' containing a set of output files. The most important ones are shown below.

summary_statistics.txt: A summary text file reports the number of genes discovered in the analyzed data, categorized into core, soft-core, shell, and cloud, based on their occurrence frequency within the studied genomes. Soft-core, shell and cloud genomes are concepts specific to Panaroo, and are a refinement of the accessory genome concept; see below for their definitions. These categories can be represented in pie charts (Figure 1A).

pan_genome_reference.fa: A multi-fasta file that contains a unique representative nucleotide sequence extracted from each cluster present in the pangenome.

gene_presence_absence.csv: This spreadsheet contains the description of each gene in the pangenome

gene_presence_absence.Rtab: A binary tab-separated matrix of presence/absence of each gene in the pangenome, the presence of a gene is coded as 1 and absence as 0.

core_gene_alignment.aln: A file that contains an alignment of all core genes (by default > 95% of samples). This can be used for phylogenetic analysis.

After constructing the pangenome, all CDSs are now clustered into orthologous groups of genes (OGs). Hereinafter OGs will be referred to simply as genes. Panaroo assigns genes to the core (present in more than 99% of genomes) or to the accessory genomes, which is subdivided into the soft-core (95–99% of genomes), shell (15–95% of genomes) and cloud (less than 15% of genomes). These statistical data of the pangenome can be represented in a pie chart and a histogram of gene frequencies. To generate these plots we will use the binary tab-separated matrix of presence/absence (gene_presence_absence.Rtab) in the following R script:

R script

```
# Read the presence/absence matrix obtained from panaroo
data <- read.table("gene_presence_absence.Rtab",sep = "\t", row.names
= 1, header = T, check.names = F)
# Calculate pangenome size
pangenome_size = nrow(data)
# Calculate the core, shell and cloud sizes as assigned by panaroo
core_size <- length(rowSums(data)[rowSums(data) > 0.99*ncol(data)])
shell_size <- length(rowSums(data)[rowSums(data) < 0.99*ncol(data) &
```

```
rowSums(data) > 0.15*ncol(data)])
cloud_size <- length(rowSums(data)[rowSums(data) <  0.15*ncol(data)])
par(mfrow = c(1, 2),pin = c(2.5, 2.5))
# Plot a pie chart displaying the core and accessory proportions; this
is Figure 1A
slices <- c(core_size,shell_size,cloud_size)
pct <- round(slices/sum(slices)*100,2)
lab <- paste(c("core", "shell", "cloud"),pct,"%",sep=" ")
pie(slices, labels = lab, main="Pangenome", cex=0.8)
# Plot a histogram of genes families in the 14 S. Typhimurium genomes;
this is Figure 1B
hist(rowSums(data), xlab = "Number of genomes containing a gene",
     ylab = "Number of genes", main = "Gene frequency",
     ylim = c(0,5000), xlim = c(0,ncol(data)+1),
     breaks = seq(min(rowSums(data))-0.5, max(rowSums(data))+0.5, by =
1))
```

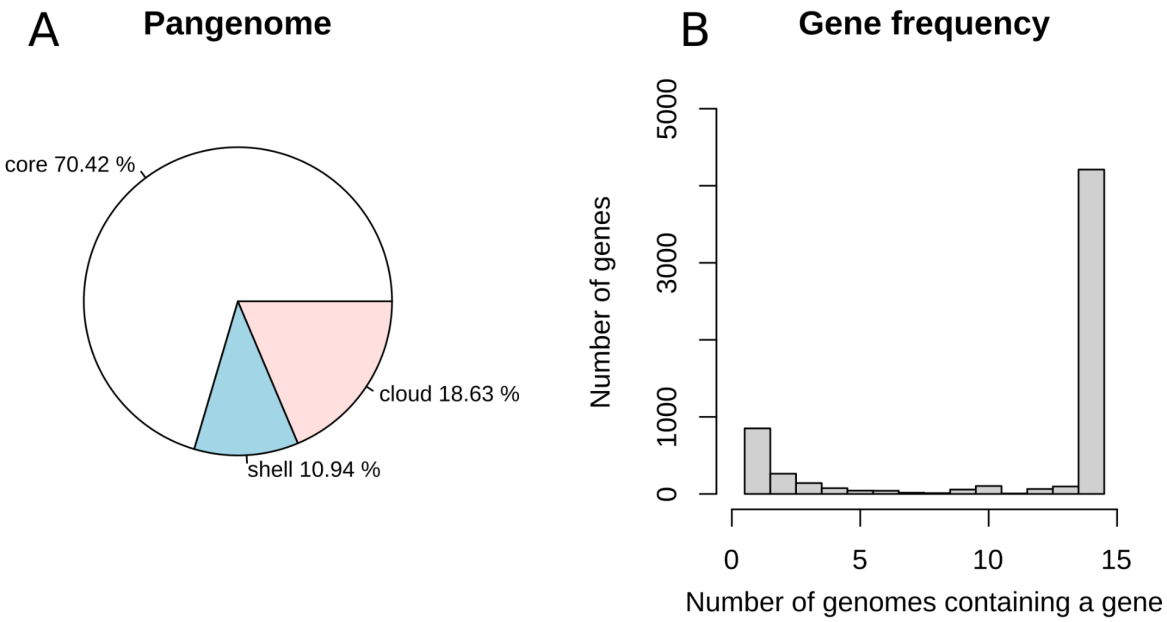

Figure 1. (A) The pie chart displays the proportion of core, shell and cloud of the SeT pangenome using 14 genomes. (B) gene frequency distribution across the number of genomes, the typical asymmetric U-shape is observed with most genes present in all genomes.

In our example, the core genome represents around 70 % (4,210 genes) of the total SeT pangenome (5,978 genes), whereas the accessory section (shell and cloud) represents about 30 % (1,768 genes). A value of 70% for the core genome is relatively high and can be explained by the fact that the genomes chosen are all strains of one serovar of the *Salmonella enterica* species; such genomes tend to share a large fraction of their protein-coding genes.

In most species, when the number of genomes analyzed increases, the size of the core genome tends to decrease, because newly added genomes may not have all the genes that are part of the previous core. The opposite happens with the accessory genome, which tends to increase with more genomes added to the analysis (however, see below for the concept of open and closed genomes). It is important to note that different pangenome reconstruction programs may yield different pangenome

estimates because they use distinct ortholog identification methods, identity cutoff values, or they may differentially account for assembly and annotation errors [53].

The gene frequency histogram (Figure 1B) displays a “U-shape” distribution, where most genes are either present in only one genome (single-genome accessory genes) or in all genomes (core genes); the intermediate-frequency accessory genes generally have lower counts. This “U-shape” distribution is typically found in prokaryote genomes and is the result of the interplay between gene loss and Horizontal Gene Transfer [4].

6.2 Open and closed pangenome

When analyzing a set of genomes using the concept of pangenome, one important question to ask is whether the pangenome for that particular set is open or closed. A pangenome is classified as ‘open’ when its always grows when new genomes are added to the computation. By contrast, a closed pangenome means that, after a certain number of genomes have been added, any new genomes for that taxon will not contain any genes not seen before.

The openness/closeness of a pangenome can be estimated by constructing rarefaction curves and applying a statistical model as proposed by Tettelin et al. [45]. A rarefaction curve is the cumulative number of unique ortholog genes we observe as more and more genomes are added to the dataset (Figure 2). The Heaps law model fits the rarefaction curve of the pangenome according to the function:

$$n = k \times N^{-\alpha}$$

Where:

- n is the expected number of genes for a given number of genomes (N),
- k and α are free parameters that are determined empirically.

According to Heaps' law (which is a power law), when α is less than 1, the pangenome is considered open, and when α is greater than 1, the pangenome is considered closed.

Rarefaction curves are computed by first performing a number of random permutations of genomes, and then fitting a power law to pangenome counts. The coefficients α and *k* can be calculated using the micropan package [28] in R with the *heaps()* function and the rarefaction curve can be computed with the *specaccum()* function of the vegan package. The script in R is shown below.

R script

```
library(micropan)
library(vegan)
# read the presence/absence matrix generated by panaroo
data <- read.table("gene_presence_absence.Rtab", sep = "\t", header =
T, row.names = 1, check.names = F)
# transpose dataframe
df_t <- t(data)
rownames(df_t) <- colnames(data)
colnames(df_t) <- rownames(data)
# calculate coefficients 'k' and 'α' setting the number of random
permutation in 1000
heap <- heaps(df_t, n.perm = 1000)
# compute rarefaction curve with 1000 permutations
rf <- specaccum(df_t, "random", permutations = 1000)
# plot the accumulation curve with the shaded area representing
confidence intervals
plot(rf, ci.type = "poly", col = "darkblue", lwd = 2,
```

```
      ci.lty = 0, ci.col = "lightblue3",
      xlab = "Number of genomes",
      ylab = "Number of gene families",
      ylim = c(4000, 6000))
# set the legend with the value of the 'α' coefficient
legend(x = "bottomright", legend = paste("\u03B1 =",
      round(heap[2], 2)), fill = "blue")
```

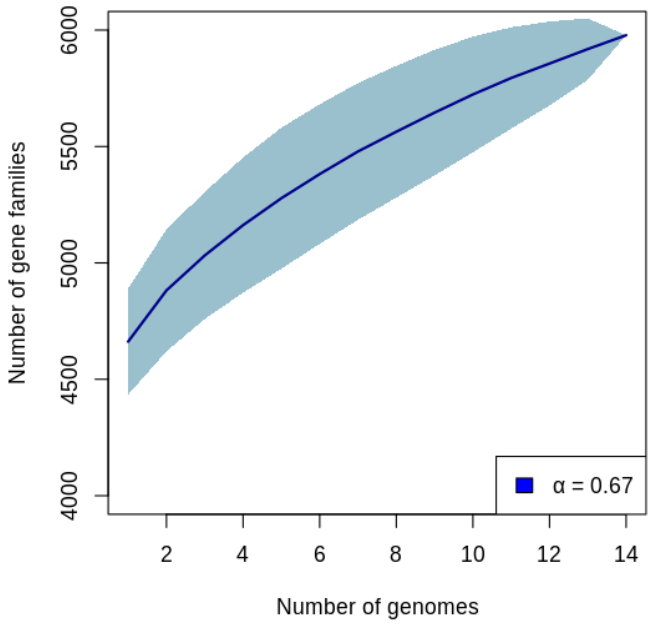

Figure 2. Rarefaction curve of the SeT pangenome calculated from random combinations of strains. The blue line indicates median values and the shadow indicates 95% confidence intervals. The legend shows the value of the parameter "α" of 0.67 indicating an open pangenome.

The SeT pangenome that we computed based on 14 genomes is open, with α = 0.67; this is consistent with results from the literature [54, 55]. However, there is also evidence that the pangenome for

Salmonella enterica is closed [56]. Conclusions regarding the pangenome openness and closeness of species can be inconsistent between studies when a small number of genomes and/or sampling bias in genome sequencing are used. Alternative metrics can quantify pangenome diversity such as genomic fluidity. This is a metric used to quantify the degree of dissimilarity in gene content between genomes. In the case of two genomes, genomic fluidity is calculated as the proportion of genes that are specific to one genome out of the total of genes present in both genomes. For a population, fluidity is determined by averaging genome fluidity calculation over all pairs of genomes [57, 58].

6.3 Comparison of gene content

Graphical representations of gene content variation using the presence/absence matrix can be depicted by Venn diagrams, presence/absence binary maps or principal component analysis (PCA).

Venn diagrams are only useful to represent gene content relationships for at most a handful of genomes. Above that, Venn diagrams become progressively more complicated. The function *vennDiagram()* of the *limma* R package can be used to draw a gene content Venn diagram as illustrated in Figure 3. The code used to plot Figure 3 from the presence/absence matrix in the R environment is as follows:

R script

```
library(limma)
# import the gene presence/absence matrix generated by panaroo
data <- read.table("gene_presence_absence.Rtab", sep = "\t", header =
T, check.names = F, row.names = 1)
# select only four genomes from the dataset
counts <- vennCounts(data[6:9])
# plot Venn diagram
```

```
vennDiagram(counts, circle.col = c("red", "blue", "green3", "yellow"),
cex =1)
```

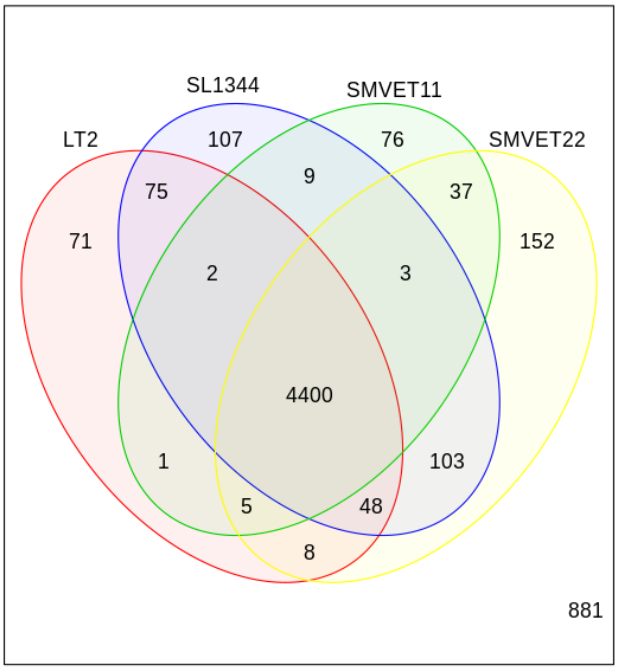

Figure 3. Illustration of a gen content Venn diagram for four SeT genomes (LT2, SMVET11, SMVET22, SL1344). The numbers inside each region indicate the number of genes shared in that region. For example, the region that represents the intersection of all four ellipses contains 4400 genes, meaning that the four genomes share these many genes. The number outside all ellipses (881) represents the number of genes that are absent in these four genomes but are present in the other 10 genomes of the dataset.

To visualize variation in the genetic content of more than a handful of genomes the presence/absence heatmap is more practical. In presence/absence heatmaps each row represents a genome, while each column represents a gene (orthologous group); the presence of a gene is denoted as a colored cell,

whereas an absent gene is represented by a non colored cell. The clustering methods to group patterns of gene presence/absence in both genomes (rows) and genes (columns) help identify genes that are unique to certain subsets of genomes or that are shared among many genomes. We used the *pheatmap* package [31] to generate a heatmap representation of the pangenome from the presence/absence matrix (FIgure 4). The script in R is shown below.

R script

```
library(pheatmap)
# read presence/absence matrix of gene content from panaroo
data <- read.table("gene_presence_absence.Rtab",sep = "\t",header = T,
row.names = 1, check.names = F)
# transpose dataframe
df_t <- t(data)
rownames(df_t) <- colnames(data)
colnames(df_t) <- rownames(data)
# convert dataframe to matrix
pange <- as.matrix(df_t, as.numeric)
rownames(pange) <- colnames(data)
# plot heatmap showed in figurer 4A
hm <- pheatmap(pange, clustering_distance_rows = "manhattan",
      clustering_method = "ward.D", color = c("white", "skyblue4"),
      clustering_distance_cols = "manhattan", show_colnames = F,
      cluster_cols = T, cluster_rows = T, legend = F)
# recover data matrix from heatmap after clustering
reorder <- data[hm$tree_col[["order"]],]
```

```
reorder <- reorder[,hm$tree_row[["order"]]]

# plot heatmap of a selected region of the pangenome matrix (figure
4B)

p2 <- pheatmap(t(reorder[5816:5890,]), show_colnames = T,

      cluster_cols = F, cluster_rows = F, legend = F,

      color = c("white", "skyblue4"))
```

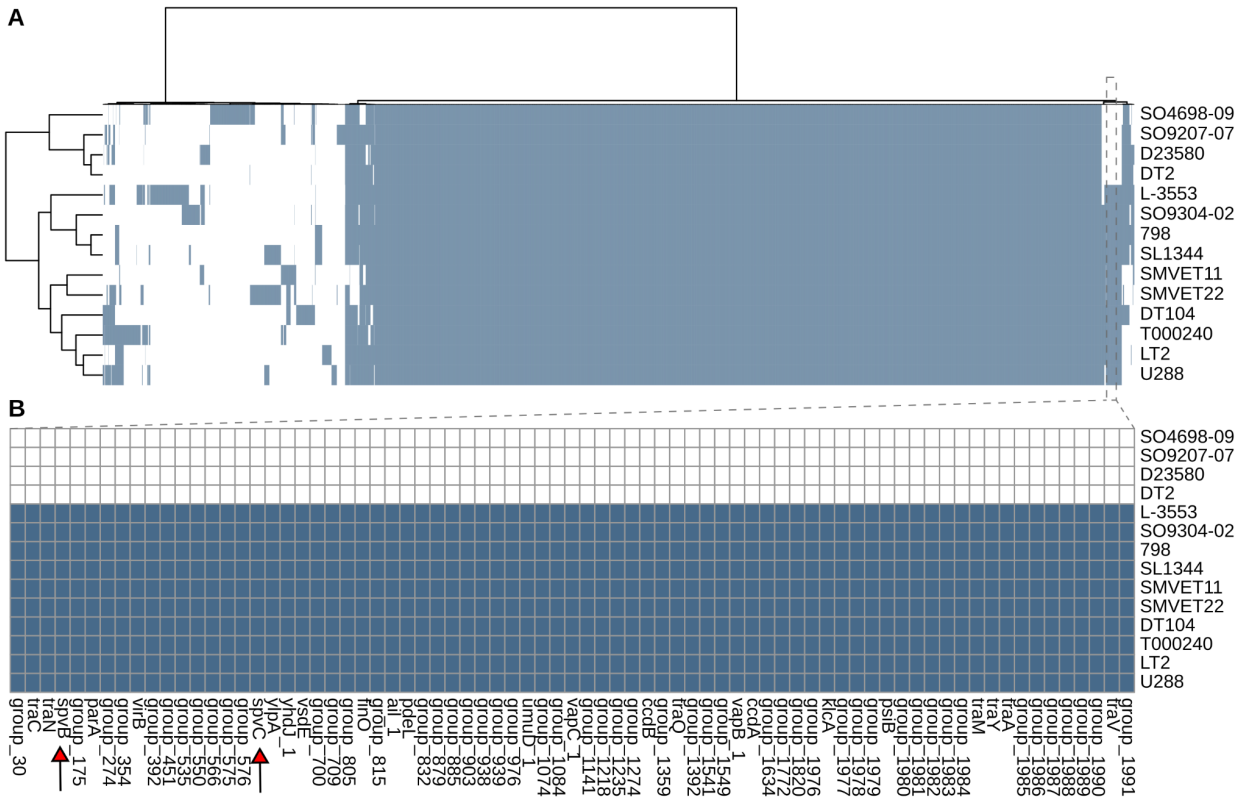

Figure 4. A. The heatmap represents the gene content, in terms of presence/absence of genes in the SeT pangenome inferred from the comparison of fourteen genomes. Gene presence is shown in blue and gene absence in white. A total of 5976 orthologous groups (genes) representing the full pangenome size are displayed. Dendrograms clustering genes (X-axis) and genomes (Y-axis) according to gene presence/absence pattern were produced in *pheatmap* package using hierarchical clustering based on the "ward.D" method with "manhattan" distance.

B. The heatmap highlights a cluster of 85 genes (dashed box in figure A) absent in four genomes (SO4698-09, SO9207-07, D23580 and DT2) but present in the other 10 genomes. These include the *spvB* ans *spvC* genes (red arrows), which are plasmid-carried virulence factors.

Figure 4A shows the full pangenome gene content with presence of a gene in blue and absent without color. The hierarchical clustering of genomes and genes allows us to observe different presence/absence patterns. For example, there is one cluster of 85 genes absent in four SeT isolates (SO4698-09, SO9207-07, D23580 and DT2) but present in the other ten genomes (figure 4B). Virulence plasmid-borne *spvB* and *spvC* genes stand out among these cluster of genes. The isolates that lack *spvB* and *spvC* genes may exhibit reduced virulence [59].

PCA is a statistical method that helps analyze genomic diversity and identify possible associations of genomes based on gene content [60]. In a PCA, the gene presence/absence matrix is first transformed into a set of principal components that capture the variation in gene content across the genomes. The principal components are then plotted in a two-dimensional space, where each point represents a genome, the position of the point reflects its gene content, and the distance between points is a measure of how different two genomes are in terms of gene content. The points can be colored according to any characteristic of the strain (e.g. host, country, sequence type). Since we have metadata associated to each SeT strain in our dataset (Table 1), we will use it to color the strains according to the host source and phylogroup variables. The R package ggplot2 can be used to perform a PCA from the gene presence/absence matrix.

R script

```
library(ggplot2)
```

```
# read the presence/absence matrix generated by panaroo
data <- read.table("gene_presence_absence.Rtab", sep = "\t", header =
T, row.names = 1, check.names = F)
# transpose dataframe
df_t <- t(data)
rownames(df_t) <- colnames(data)
colnames(df_t) <- rownames(data)
# read the metadata (table 1 of this chapter) as a dataframe
meta <- read.table("metadata.tab", header = T, sep = "\t", row.names =
1)
# merge presence/absence matrix with metadata into one dataframe
dafr <- data.frame(merge(df_t, meta, by = 0))
# compute principal components on the accessory portion of the
pangenome
PC<-prcomp(dafr[, c(4211:5979)])
# you can use "Phylogroup" instead of "Host"
PCi<-data.frame(PC$x, Host=dafr$Host)
# plot PCA labeled by Host (or Phylogroup)
ggplot(PCi, aes(x=PC1,y=PC2,fill=Host)) +
  geom_point(size = 5, alpha = 0.5, shape = 21) +
  scale_fill_brewer(palette = "Set1") +
  theme_bw()
```

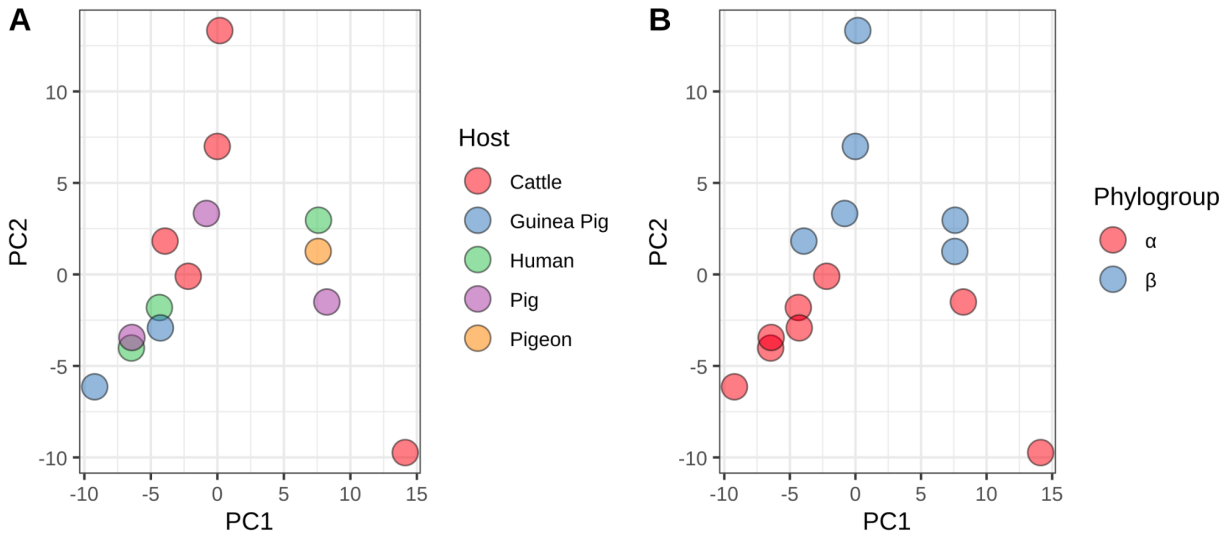

Figure 5. Principal Component Analysis of the 14 SeT genomes. The graph is generated from the gene presence/absence matrix. Each dot represents a genome, which is colored according to host source (A) or phylogroup (B). A horizontal line around y value 0 in (B) provides a good separation between the two phylogroups, suggesting that phylogroup is indeed a good determinant of gene content, and that separation is provided by the second principal component (PC2).

The PCA analysis of SeT gene content was unable to distinguish isolates from different host sources (Figure 5A). However, isolates from phylogroups α and β were separated by the second component (Figure 5B), which means that there is differential gene content between isolates from these two phylogroups.

6.4 Pangenome-wide association studies

Genome-Wide Association Study (GWAS) is an approach for studying genotype-phenotype associations. In prokaryotes, the GWAS approach is applied to pangenomes (pan-GWAS) in order to identify genes associated with specific phenotypes, such as host source, virulence, and antibiotic resistance [61]. Scoary [34], a popular tool for pan-GWAS analysis, correlates gene presence/absence from pangenome analysis with phenotypic traits. The presence/absence matrix generated by Panaroo can be used as input for Scoary.

Scoary needs two inputs: the *gene_presence_absence.csv* file generated by panaroo and a trait file (csv format) containing phenotypic traits. For example, if the trait is resistance to tetracycline and the categories are resistant and susceptible, we could use "0" to indicate susceptibility and "1" to indicate resistance.

Bash shell

```
$ scoary.py -g <gene_presence_absence.csv> -t <traits.csv>
```

Here, we cannot use our dataset because GWAS requires many more than just 14 genomes to assess the association to a specific phenotype (for additional information, see [62]). The power to find statistically significant associations is affected by several factors such as sample size, allele frequency, population diversity and structuring [63]. A large number of genomes (hundreds) are typically used in this kind of analysis [61].

7. Phylogenetic tree based on core genome alignment

Phylogeny inference is a vast topic; we include here an example of phylogeny inference because it is relatively straightforward to obtain a phylogeny given the alignment file *core_gene_alignment.aln* generated by panaroo. For more information on phylogeny inference we refer the reader to [64]. Most of the current phylogeny inference tools use a maximum likelihood approach, such as RAxML [65], fastTree [66], and IQ-TREE 2 [26].

It is important to consider the effect of homologous recombination when reconstructing phylogenies in prokaryotes. Some tools like Gubbins [25] or ClonalFrame [67] can be used to mask recombinant regions before reconstructing a phylogeny.

IQ-TREE 2 offers multiple options, including ultrafast bootstrapping (-B) and ModelFinder to find the best substitutions model (-m). The tree is typically created in Newick format and can be visualized by ggtree package [24] in R. Commands for recombinant regions detection and phylogenetic reconstruction using Gubbins and IQ-TREE 2 starting from *core_gene_alignment.aln* file generated by panaroo are shown below.

Bash shell

```
# run gubbins for detection of recombination in the alignment
$ run_gubbins -p gubbins core_gene_alignment.aln
# extract SNPs form the alignment using snp-sites
$ snp-sites -c gubbins.filtered_polymorphic_sites.fasta >
clean.core.aln
# reconstruct the phylogenetic tree with 1000 of bootstrap (-B)
$ iqtree2 -s clean.core.aln -B 1000 --prefix tree_clean_ST
```

IQ-TREE 2 takes a few minutes on a standard laptop and generates several files. The tree file (.treefile) in Newick format is used for ggtree and phangorn package in R for tree visualization.

R script

```
library(ggplot2)
library(ggtree)
library(phangorn)
# read Newick file generated by IQTREE2
```

```
tree <- read.tree("tree_clean_ST.treefile")
# set midpoint root
treeMP<-midpoint(tree)
# draw phylogenetic tree
gg <- ggtree(treeMP, layout= "rectangular", right=F) +
  geom_tiplab(size=2.8, linesize=.5,offset = 0.0003,align = T) +
  geom_text2(aes(subset = !isTip, label=label), size = 2,
             hjust=1.2,vjust = -0.3)+
  geom_treescale()
# read metadata information (table 1)
meta <- read.table("metadata.tab", header = T, sep = "\t")
# add host source information to the tree
p1 <- gg %<+% meta[,c(1:2)]
# draw tippoint colored according to host information
p2 <- p1 + geom_tippoint(aes(fill=Host, shape=Host),
                         size=5, alpha=1, colour = "black") +
  scale_fill_brewer(palette = "Set1")+
  scale_shape_manual(values = rep(21,each=12))+
  theme(legend.position = "right")
# draw vertical bars on clades representing α and β phylogroups
p3 <- p2 + geom_cladelab(node=20,barsize=1.5, label="α",
                         offset=0.0008,offset.text=.0) +
           geom_cladelab(node=19,barsize=1.5, label="β",
           offset=0.0008, offset.text=.0)
```

```
# plot
p3
```

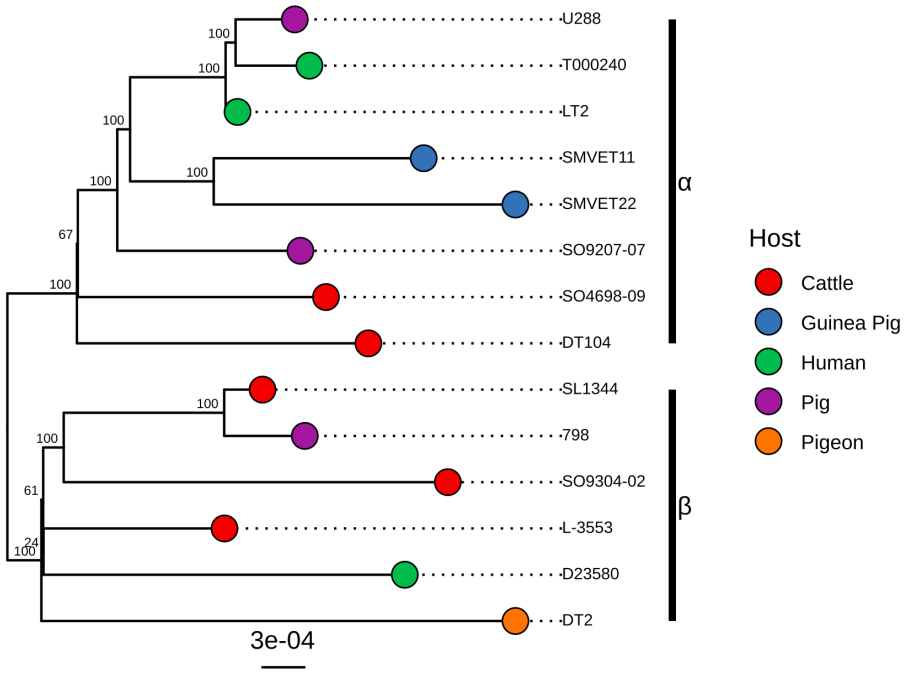

Figure 6. Phylogenetic tree of *S.* Typhimurium reconstructed from the core genome alignment of fourteen genomes using IQ-TREE 2. Tip point circle shapes are colored according to the host source. Vertical black bars represent well supported clades (α and β phylogroups).

The phylogenomic tree of SeT (Figure 6) shows two main groups, which represent the well-known α and β clades [17], both with high bootstrap support. A probable association of lineages to certain hosts is not evident due to the small number of isolates. However, previous works have shown that certain lineages are associated with some hosts, such as DT8 associated with ducks or ST313 associated with humans in Africa [68].

8. Pangenome graphs

The most common methods used to generate prokaryotic pangenomes provide a matrix indicating the presence or absence of genes, without regard for gene order or orientation [51, 69]. However, more detailed information on the variability of both gene content and genome structure within a group of genomes may provide valuable information about the evolution of the corresponding species.

Novel algorithms have emerged to address this issue, focused on extending the pangenome framework of microbial diversity to graphical models [70]. A pangenome graph is a model in which nodes represent gene families and edges represent a relation of genetic contiguity. Current tools that use this approach in prokaryotic genomes are PPANGGOLIN [30] and PanGraph [71].

In PPANGGOLIN a statistical approach is used to classify the gene presence/absence of the pangenome into persistent (gene families present in almost all genomes), shell (gene families present at intermediate frequencies), and cloud (gene families present at low frequency) partitions. PANGGOLIN uses gff/gbk annotation files as input and yields multiple output files such as gene_presence_absence.Rtab and matrix.cvs, which are also produced by panaroo. The most important output of PPANGGOLIN is an HDF-5 file named pangenome.h5. It stores all information about the pangenome, including data to construct the graph. We will now use PPANGGOLIN to compute the pangenome from the GFF files of the 14 S. Typhimurium genomes.

Bash shell

```
$ ppanggolin workflow --anno gff_list.tab
```

The gff_list.tab is a tab-separated file containing a list of the strain names and paths to the associated GFF files of all genomes used in the analysis. After the running is completed, PPANGGOLIN produces a directory containing all outputs. To visualize the partitioned pangenome graph we load the pangenomeGraph_light.gexf.gz file (this is the file that contains the graph description in terms of nodes and edges) into the Gephi program.

The Gephi software is an open source tool for exploration and visualization of networks and graphs. We use Gephi with the ForceAtlas2 algorithm and the following parameters: Scaling = 4000, Stronger Gravity = True and Gravity = 1.0. The final plot is shown in FIgure 7.

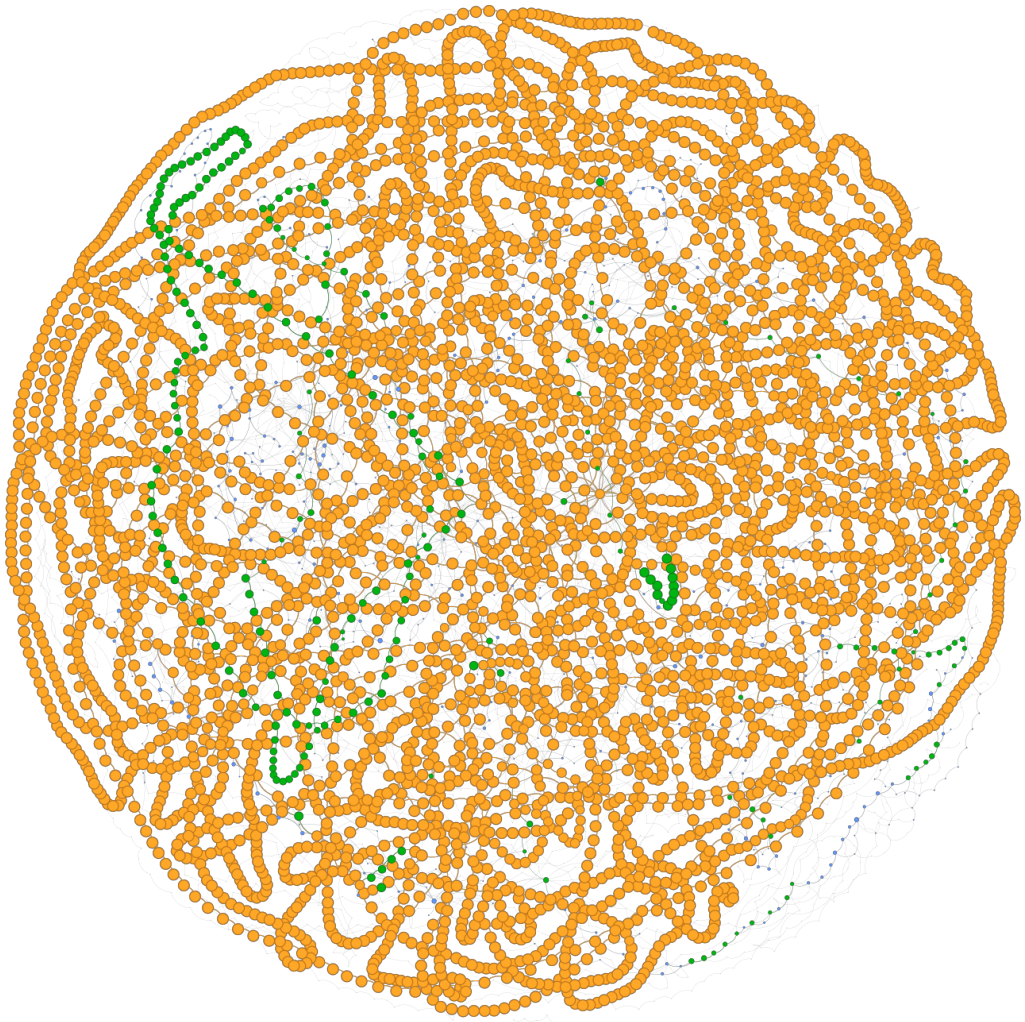

Figure 7. The partitioned pangenome graph was calculated from 14 SeT genomes. Nodes represent gene families and edges represent a relation of genetic contiguity. The persistent, shell and cloud nodes are colored in orange, green and blue, respectively. The size of the nodes is proportional to the number of strains wich share that gene.

9. Identification of Sequences of Interest in Genomic Data

9.1. Antimicrobial-resistant genes and Virulence factors prediction

Identification of Antimicrobial-Resistant Genes (ARGs) and Virulence Factors (VFs) in genomic data is standard in clinically associated bacteria. ARGs search in assembled genomes can be carried out using curated databases such as Resfinder [72], CARD [73] or NCBI-ARMfinderPlus [74], whereas VFs can be identified using the Virulence Factor Database (VFDB) [75]. The ABricate pipeline (https://github.com/tseemann/abricate) is frequently used to screen assemblies against the databases mentioned above. It runs a BLAST/DIAMOND search in FASTA files (assemblies) with customizable identity/coverage cutoffs and also allows combining reports of different runs into a matrix of gene presence/absence.

We use the genomes in fasta format downloaded from Genbank as input to run ABricate with the resfinder database as shown by the following commands.

Bash shell

```
# run ABricate with identity/coverage cutoff of 90/80 % respectively
$ abricate --db resfinder *.fasta --minid 90 --mincov 80 >
result_resfinder.tab
# combine reports into presence/absence matrix
$ abricate --summary result_resfinder.tab >
resfinder_summary.tab
```

ABricate generates an antimicrobial-resistant gene presence/absence matrix with a present gene represented by its "% of coverage" and an absent gene denoted by a point (".”). To visualize the gene presence/absence matrix in a heatmap alongside the phylogenetic tree we use the *gheatmap()* function of *ggtree* package in R with the following code:

R script

```
library(ggnewscale)
# read the gene presence/absence matrix (resfinder_total.tab)
df1 <- read.table("resfinder_summary.tab", sep = "\t", header = T,
check.names = F, row.names = 1)
# remove first row from the dataframe containing the number of ARGs
df2 <- df1[,-1]
# rename coverage values as 'P' if present and 'A' if absent
df2[df2 >= 90] <- 'Present'
df2[df2 < 90] <- 'Absent'
# start with p3 object generated in the section 7
p4 <- p3 + new_scale_fill()
# generate the gene presence/absence heatmap alongside the tree
p5 <- gheatmap(p4, df2, width = 1.2, font.size = 2.8,
      color ="black", colnames_offset_y = -0.4, colnames_position =
      "top", offset = 0.0012, hjust = 0, colnames_angle = 90) +
        scale_fill_manual(breaks = c("Present", "Absent"),
                          values = c("#6da3a3", "gray95"),
                          name = "Plasmid replicon") +
       theme(legend.position = "right")
p5 + ylim(NA,16)
```

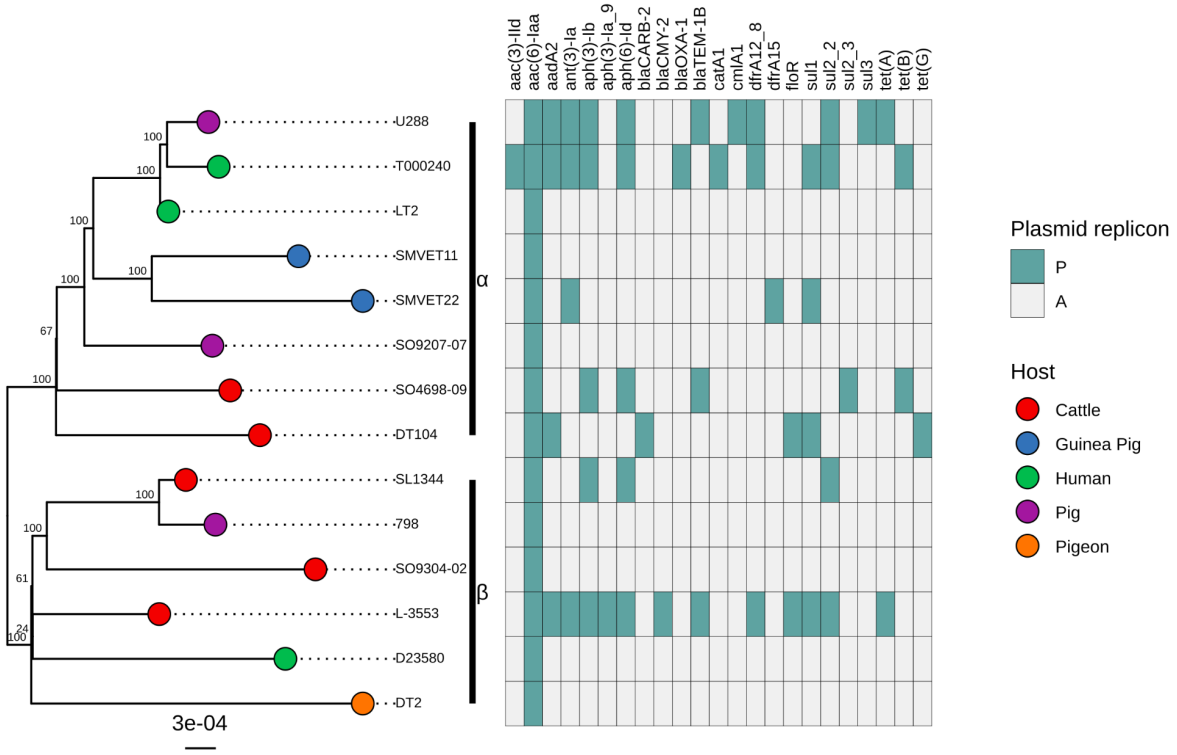

Figure 8. Phylogenetic tree of SeT coupled with a heatmap of ARGs predicted by ABRicate using the resfinder database. Turquoise boxes indicate that the gene is present; white boxes show the gene is absent.

In Figure 8 we observe that the 14 genomes we are analyzing have different ARG content. Moreover, only one ARG (*aac(6')-Iaa*) is present in all strains. *aac(6')-Iaa* gene encodes an acetyltransferase that confers resistance to aminoglycoside drugs. It was acquired before *Salmonella enterica* serotypes diversification, and is present in almost all isolates of the Typhimurium serotype [76]. Diverse ARGs were detected in humans and livestock-associated SeT strains (the columns of Figure 8). Resistance to quinolone is typically found in *Salmonella* and is due to point mutations in DNA gyrase and topoisomerase IV genes. ABricate only detects acquired resistance genes, so we recommend using pointfinder [77] to detect chromosomal mutations predictive of drug resistance.

9.2. Phage sequence prediction

Phage prediction tools are used to identify prophages and other viral sequences in assembled bacterial genomes. There are several different phage prediction tools available, and each uses different algorithms and databases. Two of the most commonly used phage prediction tools are PHASTER [78] and virsorter2 [36]. PHASTER runs on a web server and uses a combination of sequence similarity, gene prediction, and structural analysis to identify prophages in bacterial genomes. It also provides information on the location and orientation of the prophage, as well as the predicted functions of its genes. Virsorter2 is a standalone program that can be run locally, and uses a machine-learning approach to classify viral sequences in genomic and metagenomic datasets. It can identify complete or partial phage genomes, as well as other types of viruses, and it provides information on their taxonomy, gene content, and potential hosts.

To predict phage sequences for the genomes in our dataset, we loop virsorter2 through all of the genomes using the following command:

Bash shell

```
$ for i in *.fasta; do virsorter run -w "${i%%.*}"_vir -i "$i"
--include-groups "dsDNAphage,ssDNA" -j 4 --min-score 0.8
--min-length 10000; done
```

Virsorter2 produces two output files: final-viral-boundary.tsv and final-viral-boundary.fasta. The first is a tab-separated table that contains the start and end base-pair positions of all predicted phage sequences in the genome and the other file contains the sequence of these phages in FASTA format.

We can compare phage sequences and evaluate their conservation and synteny plotting comparison figures of multiple genomes. To achieve this we can use Easyfig [20], an application written

in Python with an easy-to-use graphical user interface used for creating linear comparison figures of multiple genomic loci from annotation files (e.g. GenBank).

The .gbk annotation file of LT2 and U288 genomes generated by prokka in section 5 was uploaded to Easyfig to generate a comparative plot of phage sequences. In Easyfig, set the start and end positions of the phage predicted by virsorter2. The image can be exported in SVG format.

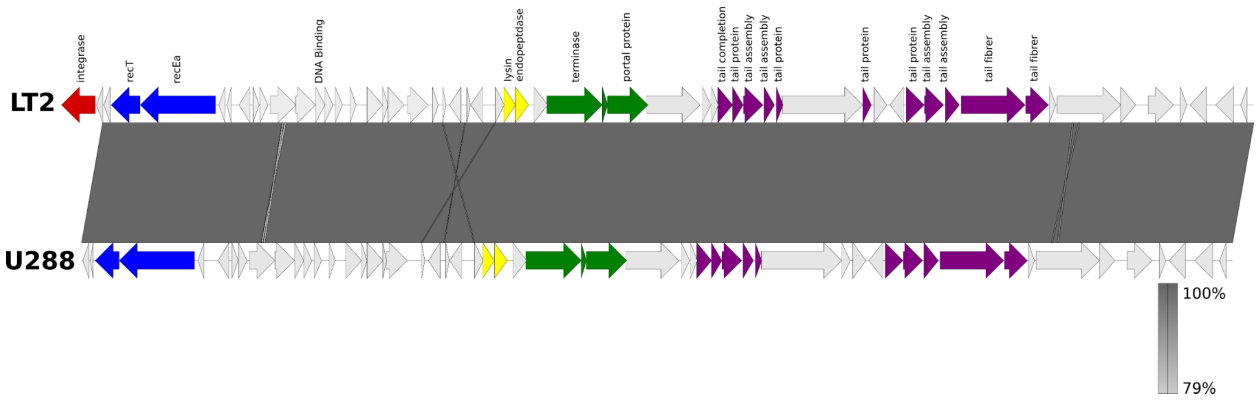

Figure 9. EasyFig output image of the phage sequences of two S. Typhimurium isolates (LT2 and U288). Coding regions are shown as arrows. Selected open reading frames are colored in relation to their functions. The percentage of sequence similarity is indicated by the intensity of the gray color.

In Figure 9, we observe sequence conservation and synteny in a phage of two SeT genomes. Previous works have revealed that phages are responsible for the main genetic content variation in *Salmonella* genomes [17, 79].

10. Conclusion

In this tutorial, we used a set of freely available software to compare fourteen SeT genomes and describe the characteristics of the pangenome, explore the genetic content variation, and perform a phylogenomic analysis, starting from genomes downloaded from GenBank. Although in our example we used a small dataset for illustrative purposes, it revealed an open pangenome for SeT of size 5,978 genes and a core genome with 4,210 genes. We identified important variations in terms of gene content, including differential presence of virulence genes between isolates; gene content distribution according to phylogroups and high synteny conservation. Additionally, we predicted ARGs in livestock- and human-associated isolates and we annotated phage sequences. The programs and codes presented here can be used by the reader interested in carrying out a comparative analysis of genomes from any bacterial species.

REFERENCES

1. Fleischmann RD, Adams MD, White O, et al (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269:496–512. https://doi.org/10.1126/science.7542800
2. Mushegian AR, Koonin EV (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc Natl Acad Sci U S A 93:10268–10273
3. Welch RA, Burland V, Plunkett G, et al (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proc Natl Acad Sci 99:17020–17024. https://doi.org/10.1073/pnas.252529799
4. Arnold BJ, Huang I-T, Hanage WP (2022) Horizontal gene transfer and adaptive evolution in bacteria. Nat Rev Microbiol 20:206–218. https://doi.org/10.1038/s41579-021-00650-4
5. Kim Y, Gu C, Kim HU, Lee SY (2020) Current status of pan-genome analysis for pathogenic bacteria. Curr Opin Biotechnol 63:54–62. https://doi.org/10.1016/j.copbio.2019.12.001
6. Ruan Z, Yu Y, Feng Y (2020) The global dissemination of bacterial infections necessitates the study of reverse genomic epidemiology. Brief Bioinform 21:741–750. https://doi.org/10.1093/bib/bbz010
7. Hurtado R, Carhuaricra D, Soares S, et al (2018) Pan-genomic approach shows insight of genetic divergence and pathogenic-adaptation of Pasteurella multocida. Gene 670:193–206. https://doi.org/10.1016/j.gene.2018.05.084
8. Mageiros L, Méric G, Bayliss SC, et al (2021) Genome evolution and the emergence of pathogenicity in avian Escherichia coli. Nat Commun 12:765.

https://doi.org/10.1038/s41467-021-20988-w
9. The CRyPTIC Consortium (2022) Genome-wide association studies of global Mycobacterium tuberculosis resistance to 13 antimicrobials in 10,228 genomes identify new resistance mechanisms. PLOS Biol 20:e3001755. https://doi.org/10.1371/journal.pbio.3001755
10. Seib KL, Zhao X, Rappuoli R (2012) Developing vaccines in the era of genomics: a decade of reverse vaccinology. Clin Microbiol Infect 18:109–116. https://doi.org/10.1111/j.1469-0691.2012.03939.x
11. Doron S, Melamed S, Ofir G, et al (2018) Systematic discovery of antiphage defense systems in the microbial pangenome. Science 359:eaar4120. https://doi.org/10.1126/science.aar4120
12. Benson DA, Cavanaugh M, Clark K, et al (2018) GenBank. Nucleic Acids Res 46:D41–D47. https://doi.org/10.1093/nar/gkx1094
13. Jolley KA, Bray JE, Maiden MCJ (2018) Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res 3:124. https://doi.org/10.12688/wellcomeopenres.14826.1
14. Markowitz VM, Chen I-MA, Palaniappan K, et al (2012) IMG: the Integrated Microbial Genomes database and comparative analysis system. Nucleic Acids Res 40:D115-122. https://doi.org/10.1093/nar/gkr1044
15. Olson RD, Assaf R, Brettin T, et al (2023) Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. Nucleic Acids Res 51:D678–D689. https://doi.org/10.1093/nar/gkac1003
16. Zhou Z, Alikhan N-F, Mohamed K, et al (2020) The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity. Genome Res 30:138–152. https://doi.org/10.1101/gr.251678.119
17. Bawn M, Alikhan N-F, Thilliez G, et al (2020) Evolution of Salmonella enterica serotype Typhimurium driven by anthropogenic selection and niche adaptation. PLOS Genet 16:e1008850. https://doi.org/10.1371/journal.pgen.1008850
18. Carhuaricra Huaman DE, Luna Espinoza LR, Rodríguez Cueva CL, et al (2022) Genomic Characterization of Salmonella Typhimurium Isolated from Guinea Pigs with Salmonellosis in Lima, Peru. Microorganisms 10:1726. https://doi.org/10.3390/microorganisms10091726
19. Seemann T (2023) ABRicate
20. Sullivan MJ, Petty NK, Beatson SA (2011) Easyfig: a genome comparison visualizer. Bioinformatics 27:1009–1010. https://doi.org/10.1093/bioinformatics/btr039
21. Cantalapiedra CP, Hernández-Plaza A, Letunic I, et al (2021) eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol 38:5825–5829. https://doi.org/10.1093/molbev/msab293
22. Bastian M, Heymann S, Jacomy M (2009) Gephi: An Open Source Software for Exploring and Manipulating Networks. Proc Int AAAI Conf Web Soc Media 3:361–362. https://doi.org/10.1609/icwsm.v3i1.13937
23. Wickham, H. (2016) ggplot2: Elegant Graphics for Data Analysis, 1st ed. Springer New York, NY
24. Yu G, Smith DK, Zhu H, et al (2017) ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol 8:28–36. https://doi.org/10.1111/2041-210X.12628
25. Croucher NJ, Page AJ, Connor TR, et al (2015) Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res 43:e15. https://doi.org/10.1093/nar/gku1196
26. Minh BQ, Schmidt HA, Chernomor O, et al (2020) IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol 37:1530–1534. https://doi.org/10.1093/molbev/msaa015

27. Smyth G, Hu Y, Ritchie M, et al (2023) limma: Linear Models for Microarray Data
28. Snipen L, Liland KH (2015) micropan: an R-package for microbial pan-genomics. BMC Bioinformatics 16:79. https://doi.org/10.1186/s12859-015-0517-0
29. Tonkin-Hill G, MacAlasdair N, Ruis C, et al (2020) Producing polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biol 21:180. https://doi.org/10.1186/s13059-020-02090-4
30. Gautreau G, Bazin A, Gachet M, et al (2020) PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. PLOS Comput Biol 16:e1007732. https://doi.org/10.1371/journal.pcbi.1007732
31. Kolde R (2019) pheatmap: Pretty Heatmaps
32. Seemann T (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153
33. R Core Team (2023) R: A Language and Environment for Statistical Computing
34. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V (2016) Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol 17:238. https://doi.org/10.1186/s13059-016-1108-8
35. Oksanen J, Simpson GL, Blanchet FG, et al (2022) vegan: Community Ecology Package
36. Guo J, Bolduc B, Zayed AA, et al (2021) VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome 9:37. https://doi.org/10.1186/s40168-020-00990-y
37. Salzberg SL (2019) Next-generation genome annotation: we still struggle to get it right. Genome Biol 20:92. https://doi.org/10.1186/s13059-019-1715-2
38. Hyatt D, Chen G-L, LoCascio PF, et al (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119. https://doi.org/10.1186/1471-2105-11-119
39. Schwengers O, Jelonek L, Dieckmann MA, et al (2021) Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. Microb Genomics 7:. https://doi.org/10.1099/mgen.0.000685
40. Hernández-Plaza A, Szklarczyk D, Botas J, et al (2023) eggNOG 6.0: enabling comparative genomics across 12 535 organisms. Nucleic Acids Res 51:D389–D394. https://doi.org/10.1093/nar/gkac1022
41. Kanehisa M, Goto S, Sato Y, et al (2014) Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res 42:D199–D205. https://doi.org/10.1093/nar/gkt1076
42. Galperin MY, Wolf YI, Makarova KS, et al (2021) COG database update: focus on microbial diversity, model organisms, and widespread pathogens. Nucleic Acids Res 49:D274–D281. https://doi.org/10.1093/nar/gkaa1018
43. Drula E, Garron M-L, Dogan S, et al (2022) The carbohydrate-active enzyme database: functions and literature. Nucleic Acids Res 50:D571–D577. https://doi.org/10.1093/nar/gkab1045
44. The Gene Ontology Consortium (2019) The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res 47:D330–D338. https://doi.org/10.1093/nar/gky1055
45. Tettelin H, Riley D, Cattuto C, Medini D (2008) Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol 11:472–477. https://doi.org/10.1016/j.mib.2008.09.006
46. McInerney JO, McNally A, O'Connell MJ (2017) Why prokaryotes have pangenomes. Nat Microbiol 2:1–5. https://doi.org/10.1038/nmicrobiol.2017.40
47. Altschul SF, Gish W, Miller W, et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2
48. Fu L, Niu B, Zhu Z, et al (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152. https://doi.org/10.1093/bioinformatics/bts565

49. Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59–60. https://doi.org/10.1038/nmeth.3176
50. Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Res 13:2178–2189. https://doi.org/10.1101/gr.1224503
51. Contreras-Moreira B, Vinuesa P (2013) GET_HOMOLOGUES, a Versatile Software Package for Scalable and Robust Microbial Pangenome Analysis. Appl Environ Microbiol 79:7696–7701. https://doi.org/10.1128/AEM.02411-13
52. Katoh K, Standley DM (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010
53. Colquhoun RM, Hall MB, Lima L, et al (2021) Pandora: nucleotide-resolution bacterial pan-genomics with reference graphs. Genome Biol 22:267. https://doi.org/10.1186/s13059-021-02473-1
54. Alikhan N-F, Zhou Z, Sergeant MJ, Achtman M (2018) A genomic overview of the population structure of Salmonella. PLoS Genet 14:e1007261. https://doi.org/10.1371/journal.pgen.1007261
55. Park S-C, Lee K, Kim YO, et al (2019) Large-Scale Genomics Reveals the Genetic Characteristics of Seven Species and Importance of Phylogenetic Distance for Estimating Pan-Genome Size. Front Microbiol 10:
56. Jacobsen A, Hendriksen RS, Aaresturp FM, et al (2011) The Salmonella enterica Pan-genome. Microb Ecol 62:487–504. https://doi.org/10.1007/s00248-011-9880-1
57. Kislyuk AO, Haegeman B, Bergman NH, Weitz JS (2011) Genomic fluidity: an integrative view of gene diversity within microbial populations. BMC Genomics 12:32. https://doi.org/10.1186/1471-2164-12-32
58. Domingo-Sananes MR, McInerney JO (2021) Mechanisms That Shape Microbial Pangenomes. Trends Microbiol 29:493–503. https://doi.org/10.1016/j.tim.2020.12.004
59. Matsui H, Bacot CM, Garlington WA, et al (2001) Virulence Plasmid-Borne spvB and spvC Genes Can Replace the 90-Kilobase Plasmid in Conferring Virulence to Salmonella enterica Serovar Typhimurium in Subcutaneously Inoculated Mice. J Bacteriol 183:4652–4658. https://doi.org/10.1128/JB.183.15.4652-4658.2001
60. Ma S, Dai Y (2011) Principal component analysis based methods in bioinformatics studies. Brief Bioinform 12:714–722. https://doi.org/10.1093/bib/bbq090
61. Allen JP, Snitkin E, Pincus NB, Hauser AR (2021) Forest and Trees: Exploring Bacterial Virulence with Genome-wide Association Studies and Machine Learning. Trends Microbiol 29:621–633. https://doi.org/10.1016/j.tim.2020.12.002
62. Didelot X (2021) Phylogenetic Methods for Genome-Wide Association Studies in Bacteria. Methods Mol Biol Clifton NJ 2242:205–220. https://doi.org/10.1007/978-1-0716-1099-2_13
63. Coll F, Gouliouris T, Bruchmann S, et al (2022) PowerBacGWAS: a computational pipeline to perform power calculations for bacterial genome-wide association studies. Commun Biol 5:1–12. https://doi.org/10.1038/s42003-022-03194-2
64. Patané JSL, Martins J, Setubal JC (2018) Phylogenomics. In: Setubal JC, Stoye J, Stadler PF (eds) Comparative Genomics: Methods and Protocols. Springer, New York, NY, pp 103–187
65. Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033
66. Price MN, Dehal PS, Arkin AP (2010) FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLOS ONE 5:e9490. https://doi.org/10.1371/journal.pone.0009490
67. Didelot X, Wilson DJ (2015) ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. PLOS Comput Biol 11:e1004041.

https://doi.org/10.1371/journal.pcbi.1004041
68. Branchu P, Bawn M, Kingsley RA (2018) Genome Variation and Molecular Epidemiology of Salmonella enterica Serovar Typhimurium Pathovariants. Infect Immun 86:e00079-18. https://doi.org/10.1128/IAI.00079-18
69. Page AJ, Cummins CA, Hunt M, et al (2015) Roary: rapid large-scale prokaryote pan genome analysis. Bioinforma Oxf Engl 31:3691–3693. https://doi.org/10.1093/bioinformatics/btv421
70. Guarracino A, Heumos S, Nahnsen S, et al (2022) ODGI: understanding pangenome graphs. Bioinformatics 38:3319–3326. https://doi.org/10.1093/bioinformatics/btac308
71. Noll N, Molari M, Neher RA (2022) PanGraph: scalable bacterial pan-genome graph construction. 2022.02.24.481757
72. Florensa AF, Kaas RS, Clausen PTLC, et al (2022) ResFinder - an open online resource for identification of antimicrobial resistance genes in next-generation sequencing data and prediction of phenotypes from genotypes. Microb Genomics 8:. https://doi.org/10.1099/mgen.0.000748
73. Alcock BP, Raphenya AR, Lau TTY, et al (2020) CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res 48:D517–D525. https://doi.org/10.1093/nar/gkz935
74. Feldgarden M, Brover V, Gonzalez-Escalona N, et al (2021) AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep 11:12728. https://doi.org/10.1038/s41598-021-91456-0
75. Liu B, Zheng D, Jin Q, et al (2019) VFDB 2019: a comparative pathogenomic platform with an interactive web interface. Nucleic Acids Res 47:D687–D692. https://doi.org/10.1093/nar/gky1080
76. Liao J, Orsi RH, Carroll LM, et al (2019) Serotype-specific evolutionary patterns of antimicrobial-resistant Salmonella enterica. BMC Evol Biol 19:132. https://doi.org/10.1186/s12862-019-1457-5
77. Zankari E, Allesøe R, Joensen KG, et al (2017) PointFinder: a novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens. J Antimicrob Chemother 72:2764–2768. https://doi.org/10.1093/jac/dkx217
78. Arndt D, Grant JR, Marcu A, et al (2016) PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res 44:W16–W21. https://doi.org/10.1093/nar/gkw387
79. Mottawea W, Duceppe M-O, Dupras AA, et al (2018) Salmonella enterica Prophage Sequence Profiles Reflect Genome Diversity and Can Be Used for High Discrimination Subtyping. Front Microbiol 9: