Summer 7-2021

The Differences of Prokaryotic Pan-genome Analysis on Complete Genomes and Simulated Metagenome-Assembled Genomes

Tang Li
University of Nebraska-Lincoln, tang.li@huskers.unl.edu

THE DIFFERENCES OF PROKARYOTIC PAN-GENOME ANALYSIS ON COMPLETE GENOMES AND SIMULATED METAGENOME-ASSEMBLED GENOMES

BY
Tang Li

A THESIS

Presented to the Faculty of
The Graduate College at the University of Nebraska
In Partial Fulfillment of Requirements
For the Degree of Master of Science

Major: Food Science & Technology

Under the Supervision of Professor Yanbin Yin

Lincoln, Nebraska

July, 2021

THE DIFFERENCES OF PROKARYOTIC PAN-GENOME ANALYSIS ON COMPLETE GENOMES AND SIMULATED METAGENOME-ASSEMBLED GENOMES

Tang Li, M. S.

University of Nebraska, 2021

Advisor: Yanbin Yin

Metagenomic assembly is often used in microbiome research. In metagenomic assembly, contigs are binned based on the shared nucleotide composition. These contig bins are called metagenome-assembled genomes (MAGs), each representing a unique bacterial genome recovered from metagenome sequencing. Hundreds of thousands of high-quality MAGs of various ecological environments have been published since 2017, and increasingly more MAGs are being used in pan-genome analyses where unculturable species or species without reference genomes are studied in microbiome research. However, compared to the traditional pan-genome analysis that uses isolate genomes (from a pure strain isolated from a mixed bacterial population), it is not known how the quality of pan-genome analyses is affected by the problems often associated with MAGs, such as fragmentation, incompleteness, and contamination. The purpose of this study is to investigate differences in pan-genome analysis on complete isolate genomes and MAGs. The specific aims are: (1) to evaluate the changes in the core genome of MAGs, and (2) to test the influence of MAGs on downstream functional analysis. MAGs with expected quality were simulated from complete genomes of 17 prokaryotic species, and pan-genome analysis was performed for simulated MAGs to generate core genomes.

Functional and phylogenetic analyses were performed using the results of simulated MAGs and benchmarked against those using the original complete genomes. Compared to the analyses using the complete genomes, fragmentation and incompleteness in MAGs led to reduced core genomes, while contamination in MAGs resulted in large numbers of unique genes. The potential underestimation in functional prediction and incorrect phylogenetic reconstruction was associated with the loss of core genomes. We suggest that more relaxed parameters should be used in pan-genome analysis to improve the accuracy on MAGs. Better quality control of MAGs and the development of new pan-genome analysis tools (e.g., with improved gene annotation and clustering algorithms) are needed in future studies.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor Dr. Yanbin Yin for the guidance and encouragement during my master's program. I am grateful for your constant technical support and spiritual encouragement, and also appreciate your patience and invaluable mentoring in research and thesis writing. I will always remember the precious experiences in research and life you shared with me. I would also like to thank my committee members, Dr. Etsuko Moriyama and Dr. Byron Chaves, for the constructive suggestions and comments on my research. Dr. Moriyama provided selfless help and instructions in my improvements in scientific writing and critical thinking. Dr. Chaves supported me with inspiring and critical comments.

I owe many thanks to my parents for their endless love and support. They taught me to believe in myself with their praise and trust. I would like to thank Dr. Long Chen for his accompany and support. Long gave me great support in research and life, and also brought happiness to me. I also thank Yafan Yu and Yiyi Cheng for always being with me. I will never forget the precious memories we had. I appreciate all the generous help from members in Dr. Yin's lab, Dr. Xuehuan Feng, Dr. Jinfang Zheng, Bowen Yang, and Yuchen Yan, and the invaluable friendships with them.

Finally, I appreciate the research fellowship provided by the Department of Food Science and Technology at the University of Nebraska-Lincoln. I am also thankful for the help from everyone in the Department of Food Science and Technology and Nebraska Food for Health Center.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

LIST OF APPENDICES

CHAPTER I

INTRODUCTION

The term pan-genome was first introduced in 2005. The pan-genome represents the entire gene set of all strains in a species (Tettelin *et al.*, 2005), which contains the core genome and the variable genome. The core genome is the set of homologous genes that are present in all genomes. Historically, complete isolate genomes of bacteria have been used for pan-genome analysis. More recently, pan-genome analyses have been increasingly used to study metagenome-assembled genomes (MAGs) of the unculturable species or species without reference genomes. MAGs are produced through DNA sequencing, read assembly, and contig binning of environmental samples, and can be regarded as a close representation of individual genomes. However, compared to fully assembled genomes, MAGs are known to have three major limitations: fragmentation, incompleteness, and contamination. These limitations may affect the accuracy of pan-genome analysis. To understand the importance of the quality of MAGs, in this study both individual genomes and simulated MAGs were used for multiple pan-genome analyses and the results were compared.

The literature review in chapter II provides background information about pan-genome and its applications, examines challenges in pan-genomic computational tools, and introduces the nature of MAGs. The overview and discussion of recent studies using MAGs in the pan-genome analysis are also included in this chapter.

The materials and methods are provided in chapter III. In this chapter, the data collection and literature review are described, and the bioinformatics pipeline to generate

simulated MAGs from complete genomes is introduced. The detailed information about pan-genome analysis and downstream functional analysis is also explained.

In chapter IV, the main results of this study are described in detail. Fragmentation and incompleteness of MAGs were found to lead to the loss of the core genome. The finding was consistent irrespective of the pan-genome tools (Roary, BPGA, Anvi'o) used. However, inconsistent results on contamination of MAGs were observed between Roary and the other two tools. Different parameters in the pan-genome analysis were also tested. It showed the importance of parameter selection for more reliable pan-genome analysis with MAGs. It also revealed the bias or errors in functional and phylogenetic analyses for MAGs.

The discussion and conclusion are provided in the last chapter. Expected as well as unexpected results are discussed. The loss of core genomes in MAGs revealed that improvement is needed in MAG quality control and gene annotation and clustering algorithms in pan-genome analysis. Finally, suggestions for pan-genome parameter selection were provided based on the quality of MAGs.

CHAPTER II

LITERATURE REVIEW

Pan-genome Definitions and Applications

The genes in a pan-genome are classified into three categories: (i) core genes, (ii) dispensable, accessory, or flexible genes, and (iii) unique/singleton genes. The core genes are shared by all strains within the species, while accessory genes present in a subset of the strains, and the unique genes are only present in a specific strain (Tettelin *et al.*, 2005). Core genes are likely to be essential for the growth or survival of any strains of the species (Tettelin *et al.*, 2005; Kim *et al.*, 2020), while the accessory and unique genes may reflect the evolutionary innovation and adaptations of each strain to its particular environment, host, and/or lifestyle (Tettelin *et al.*, 2005; Zhang and Sievert, 2014; Nelson and Stegen, 2015). In some cases, the core genes also include more conserved accessory genes and can be separated into strict-core and soft-core genes. Strict-core genes are found in >99% (Page *et al.*, 2015) or 100% (Kaas *et al.*, 2012) of the genomes studied, while soft-core genes are present in >95% (Bezuidt *et al.*, 2016) or >96% (Laing *et al.*, 2017) of the genomes studied. The inclusion of soft core genes is important when draft genomes are included in the pan-genome analysis, as some genes may be missing in draft genomes due to their lower assembly quality than completely assembled genomes (Nelson and Stegen, 2015).

The result of pan-genome analysis depends on how many genomes are included in the analysis. Pan-genomes of different species can be defined as “open” or “closed” based on power law or Heaps’ law model (Tettelin *et al.*, 2008). In other words, the

classification of the pan-genomes depends on the probability of finding new gene families as new genomes are sequenced and added into the analysis (Costa *et al.*, 2020). If the size of the pan-genome increases constantly with the addition of new genomes, the pan-genome is considered open. The pan-genome is defined as closed when the addition of new genomes does not change the size of the pan-genome significantly (Tettelin *et al.*, 2008; Carlos Guimaraes *et al.*, 2015). The open or closed nature of a pangenome is correlated with the lifestyle of the bacterial species (Medini *et al.*, 2005). Sympatric species, which live in a community and frequently contact with other species in the same environment, tend to have open pangenomes. These species tend to have a higher rate of horizontal gene transfer to continuously gain new genes. In contrast, the allopatric species that live in an isolated environment usually have a smaller and closed pan-genome (Medini *et al.*, 2005; Georgiades and Raoult, 2011).

Pan-genome analysis can be used in various studies of prokaryotic species, including genomic diversity characterization, bacterial evolutionary analysis, disease outbreak analysis, the study of virulence-associated genes, and the antimicrobial resistance study (Anani *et al.*, 2020; Kim *et al.*, 2020). In the past, pan-genome analyses have been conducted to reveal the genomic diversity and phylogeny of various species such as *Escherichia coli* (Vieira *et al.*, 2011; Kaas *et al.*, 2012), *Staphylococcus epidermidis* (Conlan *et al.*, 2012), *Bifidobacterium longum* (O'Callaghan *et al.*, 2015), *Salmonella enterica* (Laing *et al.*, 2017), and *Corallococcus spp.* (Livingstone *et al.*, 2018). Other examples include the study of an outbreak of *Staphylococcus aureus* by a pan-genome analysis of isolates from patients (Roisin *et al.*, 2016), and the identification of diagnostic marker genes in *Campylobacter jejuni* strains (Buchanan *et al.*, 2017). The

core genome of *Acinetobacter baumannii* was used to determine its carbapenem resistance (Higgins *et al.*, 2017). The *Pseudomonas aeruginosa* pan-genome provided new insights on its antimicrobial resistance and virulence genes (Freschi *et al.*, 2019).

Pan-genome Computational Tools and Their Challenges

A great number of computational tools/packages/pipelines for pan-genome analysis have been developed in the last 15 years, such as PGAP (Zhao *et al.*, 2012), GET_HOMOLOGUES (Contreras-Moreira and Vinuesa, 2013), ITEP (Benedict *et al.*, 2014), Roary (Page *et al.*, 2015), Anvi'o (Eren *et al.*, 2015), BPGA (Chaudhari *et al.*, 2016), and PanX (Ding *et al.*, 2018). A recent survey identified more than 40 pan-genome analysis tools currently available online or downloadable (Anani *et al.*, 2020). Computational tools for pan-genome analysis have been compared in previous studies (Marschall *et al.*, 2018; Anani *et al.*, 2020; Kim *et al.*, 2020; Vernikos, 2020; Pantoja *et al.*, 2020). In these studies, the authors summarized the features provided in each tool, evaluated the performance and computational efficiency, pointed out unaddressed issues and challenges, and provided further suggestions for developing better tools.

For example, Marschall *et al.* (2018) suggested that one of the important considerations in the pan-genome analysis is the "completeness" so that all functional elements should be included in a sufficient number of genomes. Bonnici *et al.* (2021) summarized that a general pan-genomic workflow includes three steps: (i) detection and annotation of genes in unannotated genomic sequences (assembled genomes or contigs), (ii) clustering homologous genes based on gene similarities at sequence or functional level, and (iii) presenting pan-genomic matrix showing the relationships between gene

clusters and the genomes they belong to. Using several benchmark datasets generated by varying the model parameters to simulate variation in gene abundance and alteration in sequences, they found that the performance of pan-genome computational tools was influenced by the input genome compositions. The performance of pan-genome computational tools decreased with increasing levels of genome variations, evolutionary distance, and the number of genomes (Bonnici *et al.*, 2021). Wu *et al*. (2021) used a series of *Bacillus subtilis* strain sets to understand the impacts of various confounding strains on the output accuracy of pan-genome analyses. They concluded that the performance of pan-genome analyses was influenced by the inclusion of incorrectly classified strains, phylogenetically distinct strains, genome-reduced/chimeric strains, strains that contain a large number of unique genes or pseudogenes, and multiple clone strains. They suggested that the quality control of input genomes was extremely important to improve the efficiency and accuracy of pan-genome analysis (Wu *et al.*, 2021). Lastly, Zhou *et al*. (2020) mentioned that the incompleteness and inconsistent gene annotations (for instance, fragmented genes missed in functional prediction) may affect the results of pan-genome analysis and lead to reduced core genome size and overestimated pan-genome size. If the orthologous genes (evolved by vertical descent) and paralogous genes (derived from gene duplications or horizontal gene transfer events) were not correctly differentiated, the inclusion of paralogous genes in the phylogenetic analysis may also cause inaccurate biological interpretation (Zhou *et al.*, 2020). All in all, the quality of the input genomes and the algorithm used in the computational tools are both important for pan-genomic analysis.

Metagenome-Assembled Genomes (MAGs) and Limitations

The term metagenomics was first proposed by Jo Handelsman in 1998 (Handelsman *et al.*, 1998). It is a culture-independent analysis that uses DNA sequencing techniques to study the genomes recovered directly from environmental samples (also known as metagenomes) (Riesenfeld *et al.*, 2004). Metagenomic studies included environmental DNA cloning, functional expression screening (Handelsman *et al.*, 1998), random shotgun sequencing, and reconstruction of environmental DNA (Venter *et al.*, 2004; Tyson *et al.*, 2004). Compared to the traditional culture-dependent method, metagenomics can be used to study unculturable microorganisms (Riesenfeld *et al.*, 2004; Tyson *et al.*, 2004; Taylor-Brown *et al.*, 2017), discover diversity and abundance of community members (Schloss and Handelsman, 2008; Delmont *et al.*, 2011; Saghaï *et al.*, 2015), and explore the metabolic potential of the community and its members (Martin-Cuadrado *et al.*, 2008; Simon *et al.*, 2009).

Metagenome-assembled genomes (MAGs) are produced from metagenome shotgun sequencing reads through filtering, assembling, binning, and taxonomy assignment steps to generate the close representation of actual individual genomes. The term MAG first appeared in the literature in 2015 (Hugerth *et al.*, 2015; Garcia *et al.*, 2015), although the earliest use of metagenome assembly and binning to recover individual genomes can be dated back to 2004 (Tyson *et al.*, 2004; Venter *et al.*, 2004). Thanks to the development of faster and more accurate contig binning tools (e.g., CONCOCT (Alneberg *et al.*, 2014), MaxBin (Wu *et al.*, 2014), ABAWACA (Brown *et al.*, 2015), and MetaBAT (Kang *et al.*, 2015)), the first large scale MAG study was published in 2015 (Brown *et al.*, 2015), although the term genome bin instead of MAG

was used. In 2016, a review paper summarized the workflow of recovering MAGs from metagenome sequencing (Sangwan *et al.*, 2016). In 2017, the Genomic Standards Consortium (GSC) published the Minimum Information about a Metagenome-Assembled Genome (MIMAG), a metagenomics community standard for publishing MAGs with mandatory metrics (genome completeness and contamination) (Bowers *et al.*, 2017). These community-driven efforts have significantly boosted the use of MAGs in large-scale microbiome research. A Google Scholar search found only 47 records with the term "Metagenome-assembled genome" before 2017, but 1,190 records after 2017. Indeed since 2017, hundreds of thousands of MAGs have been reconstructed from the various environments, including the ocean (Tully *et al.*, 2018; Jégousse *et al.*, 2021), soil (Kroeger *et al.*, 2018), lakes (Vavourakis *et al.*, 2018), the human gut (Almeida *et al.*, 2019; Pasolli *et al.*, 2019), activated sludges (Singleton *et al.*, 2021), and the animal gut (Chen *et al.*, 2021; Peng *et al.*, 2021; Watson, 2021). These MAGs have been used to (i) identify complete genes and operons to improve predictions of metabolic capacities, (ii) provide information about gene synteny and enable better taxonomic profiling, (iii) discover completely novel taxa, (iv) find and study genes in specific species/strains, (v) construct the new tree of life, and (vi) compare the abundance of different MAGs across samples (Quince *et al.*, 2017; Chen *et al.*, 2020; Bharti and Grimm, 2021).

However, there are some concerns about the use of MAGs. Chen *et al.* (2020) concluded that factors including gaps, assembly errors, chimeras, and contamination would significantly limit the advantages of using MAGs. For example, MAGs may have some gaps due to the low coverage of short reads (Chu *et al.*, 2019). MAGs may have assembly fragmentation caused by strain divergence. Therefore, assembly of long

fragments from short reads is difficult when within-population diversity is high (Chen *et al.*, 2020). Different types of assembly errors including repeat collapse, insertions, deletions, and inversions, may be involved in MAG assembly and thus influence the quality of MAGs (Olson *et al.*, 2019). The chimeric reads result in a fragmented assembly by introducing an erroneous assembly graph, leading to truncated contigs (Alneberg *et al.*, 2018). The chimeras of sequences from two different organisms may be created in misassembly (Mineeva *et al.*, 2020). Additionally, if the scaffolds are too short, the binning will produce unreliable MAGs due to binning errors. MAGs may be contaminated by phage or plasmid genome fragments that having similar coverage or GC content (Chen *et al.*, 2020).

According to the minimum information about a metagenome-assembled genome (MIMAG) mentioned above (Bowers *et al.*, 2017), MAGs are considered as "high-quality" if they are >90% complete with less than 5% contamination (Bowers *et al.*, 2017). MAGs with low completeness or high contamination can lead to incorrect conclusions. Therefore, completeness and contamination are two important metrics that need to be reported for new MAGs to determine their quality. CheckM is one of the most widely used tools that estimate both completeness and contamination in MAGs based on the presence of single-copy genes (SCGs) (Parks *et al.*, 2015). However, two partial genome bins of different genomes may be incorrectly combined (Parks *et al.*, 2015), which will limit the value of MAGs (Becraft *et al.*, 2017). Furthermore, the absence of multiple copies of SCGs cannot indicate the absence of fragments from other organisms (Chen *et al.*, 2020), CheckM may fail to detect contamination from lineages that are not represented in the database and significantly underestimated the real contamination in

MAGs (Becraft *et al.*, 2017). Although some tools like Anvi'o (Eren *et al.*, 2015) enable the manual curation of contamination beyond the use of SCGs for estimation, these strategies have limited scalability from large-scale datasets or samples. Overall, compared to complete isolate genomes, MAGs may suffer from fragmentation, incompleteness, and contamination. The quality evaluation of MAGs is critical to give a precise and meaningful interpretation in microbiology and environmental genomics.

Using MAGs in Pan-genome Study

In the past five years, the pan-genome analyses of MAGs have been increasingly used to study microbiomes in a variety of environments. One of the earliest studies was published in 2017 (Anderson *et al.*, 2017), where MAGs were reconstructed from two hydrothermal vents to investigate the genomic variations within subseafloor microbial populations; the Integrated Toolkit for the Exploration of microbial Pangenomes (ITEP) (Benedict *et al.*, 2014) was used to generate the clusters of open reading frames (ORFs) to determine the functional enrichment in *Sulfurovum* MAGs. In another study, Meyer *et al.* (2017) performed pan-genome analysis for *Roseofilum* MAGs to explore the functional repertoire of the black band disease (BBD) consortium. It was revealed that the resistance to sulfide was an important characteristic for the growth and survival of the BBD consortium.

More recently, pan-genome analyses of MAGs have been used to explore the human microbiomes in the intestinal tract, mouth, skin, and vagina. For example, a pan-genome analysis of human MAGs was performed to study the human-associated microbial diversity in different human populations (Pasolli *et al.*, 2019). Almeida *et al*.

built the Unified Human Gastrointestinal Genome (UHGG) collection by combining MAGs and isolate genomes from various resources and reducing redundancy (Almeida *et al.*, 2021). The pan-genome of each UHGG species was further generated to study intraspecies genomic diversity and the functions encoded by the core and accessory genes (Almeida *et al.*, 2021). In addition, the pan-genome of MAGs from newly identified species-level operational taxonomic units (OTUs) was constructed to study the phylogenetic diversity of newly sequenced gut bacteria (Nayfach *et al.*, 2019).

There are also pan-genome analyses of MAGs in specific species. The differences in pan-genome sizes of four *Prevotella copri* clades were compared between individuals on non-westernized diets and westernized diets. The findings revealed that *P. copri* has substantial genomic and functional diversity that were underrepresented in western-lifestyle populations (Tett *et al.*, 2019). Another report has studied the pan-genomes and core genomes of *Faecalibacterium*-like species-level genome bins (SGBs, equivalent to MAGs of the same species). It was found that the higher diversity of SGBs may be associated with increased utilization of plant-based foods, while the lower diversity of SGBs observed in western populations may be related to intestinal inflammatory and obesity (De Filippis *et al.*, 2020). Baker *et al.* (2021) conducted a pan-genome analysis of oral MAGs, which not only increased the diversity of oral species but also illustrated the significant variations in functional potential among different *Saccharibacteria* clades. Zhou *et al.* (2018) studied the origin of *Salmonella enterica* Paratyphi C by combining MAGs reconstructed from metagenomes of an old skeleton with genomes of modern *S. enterica*. The pan-genome of oral MAGs has also been used to identify shared/unique

genes and functions to determine functional markers of niche specificity (Shaiber *et al.*, 2020), and reveal associations between species and specific habitats (Utter *et al.*, 2020).

The pan-genome analysis of MAGs has also been conducted on microbiomes in the ocean, thermal vents, lakes/rivers, and soil. Pan-genome analysis revealed genomic stability and environmental adaption of ammonia-oxidizing archaea in the deep ocean (Wang *et al.*, 2019). Wilkins *et al*. (2019) performed a comparative study to identify shared genetic contents among bacteria and archaea species in two hot springs to understand the phylogenetic diversity and functional potential within springs. Moulana *et al.* (2020) found that important factors like nutrient limitation may drive the adaptation and evolution of microbial lineages in hydrothermal vents. Sheridan *et al.* (2020) determined the influence of gene duplication on the evolution and genome expansion of archaea lineages in rivers. The pan-genome of soil MAGs revealed core gene clusters and their functions in carbohydrate metabolism, and provided new biological insights for soil microbial communities (Kroeger *et al.*, 2018).

The pan-genome analysis is often followed by a more in-depth downstream analysis. The core genes identified in MAG pan-genome analysis can be used for phylogenetic analyses (Pasolli *et al.*, 2019) or serve as effective taxonomic marker genes for species identification (Nayfach *et al.*, 2019). The functional predictions generated from MAG pan-genome analysis can be applied to developing new culturing strategies for species isolation (Almeida *et al.*, 2021). Functional enrichment analysis in the core/unique genes of MAGs can shed new insights on the species evolution and adaptation to a specific environment (Moulana *et al.*, 2020; Shaiber *et al.*, 2020).

In summary, the pan-genome analysis that combines MAGs with reference isolate genomes has been used routinely to study the genomic diversity and population structure of environmental microbiomes (Reveillaud *et al.*, 2019). However, due to the nature of MAGs (fragmentation, incompleteness, and contamination), to which extent the accuracy of pan-genome results is influenced by the quality of MAGs has not been determined in previous studies. Our hypothesis is that given the quality of MAGs, which is not as good as completely assembled individual genomes, there will be biases and errors in the MAG pan-genome analysis.

CHAPTER III

METHODOLOGY

Literature Search and Summary

In order to find relevant literature that has used pan-genome analysis on MAGs, a keyword search was performed against PubMed and Google Scholar using the following queries:

(i). Search in PubMed: ((pangenome) OR (pan-genome)) AND ((metagenome-assembled genomes) OR (MAGs)),

(ii). Search in Google Scholar: ("pangenome" OR "pan-genome") AND ("metagenome-assembled genome" OR "MAGs")

The publications collected were further filtered to keep those meeting the following criteria: (i) organism(s) used in the study were from prokaryote (bacteria/archaea), (ii) at least one metagenome-assembled genome reconstructed from metagenome samples were involved in pan-genome analysis, and (iii) the specific pan-genome computational tools were listed.

All the publications that met the requirements were downloaded and manually curated. Information including PMID, title, authors, year of publication, journal, source/habitat of metagenomes, computational tools used in the pan-genome analysis, computational tool parameters, and downstream analysis was recorded (Appendix A).

The bar graphs for the year of publication and the pie chart for the source of metagenomes were created by using R ggplot2 (Wickham, 2016).

Data Collection

The bacteria assembly summary file was downloaded from the NCBI RefSeq database (ftp.ncbi.nih.gov:/genomes/refseq/bacteria/assembly_summary.txt) in October 2019 (O'Leary *et al.*, 2016). A total of 17 species, each containing at least 100 complete genomes (without gaps in the genomes), were selected for benchmarking data. All complete genomes of the 17 species were downloaded in nucleotide fasta format (the total number of downloaded complete genomes was 4,795). To filter out misannotated genomes, for each species, all-against-all average nucleotide identity (ANI) among complete genomes were calculated by using FastANI v1.32 (Jain *et al.*, 2018). Genomes with <90% of their pairwise ANI values >94% (Konstantinidis and Tiedje, 2005; Richter and Rosselló-Móra, 2009) were removed. All the remaining complete genomes of a species were used as the species dataset (e.g., *E. coli* species dataset).

To create contaminated datasets (see below), the genus-level (interspecies) contamination datasets were collected for four selected species, namely *Burkholderia pseudomallei*, *Bacillus subtilis*, *Klebsiella pneumoniae*, and *Streptococcus pyogenes*. For each species, genomes from other species within the same genus were downloaded. A total of 1,589 and 1,118 genomes were used as genus-level contamination datasets for *B. pseudomallei* and *K. pneumoniae*, respectively. One thousand genomes randomly selected from 4,189 genomes in Bacillus genus and 1,000 genomes randomly selected from 12,533 genomes in Streptococcus genus were used as the genus-level contamination datasets for *B. subtilis* and *S. pyogenes*, respectively.

Assessment of 17 species

For each species, the number of chromosomes and plasmids in each of the complete genomes were calculated. The maximum, minimum, and average number of chromosomes and plasmids in all complete genomes in the species were summarized.

To depict the phylogenetic relationship among the 17 species, one representative genome of each species was selected. The genome in each species dataset that was labeled as the "reference genome" or "representative genome" in the bacteria assembly summary was selected as the representative genome. To reconstruct a phylogenetic tree, the orthogroups in the 17 representative genomes were determined by using OrthoFinder v2.52 (Emms and Kelly, 2019). All single-copy orthogroups (the orthogroup with exactly one gene from each species) were extracted. Multiple sequence alignment (MSA) for each single-copy orthogroup was created with Muscle v3.8 (Edgar, 2004). All single-copy orthogroup MSAs were combined and the conserved blocks were determined by using gblocks v0.91b (Castresana, 2000). The phylogenetic trees were built by using RAxML v8.2.12 (Stamatakis, 2014) and visualized by using Interactive Tree Of Life (iTOL) v5 (Letunic and Bork, 2021).

Pan-genome Analysis for Each Species Dataset

For each species dataset, pan-genome analysis was performed by using all complete genomes in the dataset. All nucleotide sequences were annotated by using the automatic pipeline Prokka v1.13 (Seemann, 2014), in which the program Prodigal (Hyatt *et al.*, 2010) was used for gene finding and translation. The average genome size (the number of genes) of species complete genomes was calculated. Gene functions were

predicted in Prokka by using translated protein sequences as queries to search against a set of public databases (Seemann, 2014). The general feature format (.gff) files for genomes were used as inputs to pan-genome analysis by using Roary v3.13 (Page *et al.*, 2015) with the parameters '-i 90' (minimum amino acid identity of 90% for a positive match in blastp), '-cd 100' (a core gene defined as 100% presence), '-s' (do not split paralogs), and '-e -n' (create fast core gene alignment with MAFFT (Katoh and Standley, 2013)).

The Heaps' law model ($n=\kappa N^{\gamma}$, where n is the pan-genome size, N is the number of genomes used, and κ and γ are fitting parameters) was used in the pan-genome of each species dataset to predict the openness and closeness of the pan-genome (Tettelin *et al.*, 2008; Park *et al.*, 2019). If $\gamma > 0$, the pan-genome was considered to be open; otherwise, the pan-genome was considered to be closed. In addition, the Pearson's correlation tests were performed using cor.test function in R to determine the correlations between the number of genomes/the average genome size and the pan-genome size.

Statistical Analysis of MAGs

The 276,349 UHGG MAGs reconstructed from human gut metagenomes in the study by Almeida *et al.* (2021) were used to determine the distribution of contig number, completeness, and contamination in MAGs. A summary file containing the genome metadata was downloaded from the MGnify FTP site (ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes/human-gut/v1.0/genomes-all_metadata.tsv). The average number of contigs, completeness, and contamination of all MAGs from all species were computed and plotted. Similar plots were also made for the

MAGs from a specific species of the 17 species (e.g., *Escherichia coli* and *Klebsiella pneumoniae*). Histograms and density plots for showing the statistical analysis results were created using R ggplot2 (Wickham, 2016)

Simulate MAGs from Complete Genomes

For each species, 100 complete genomes were randomly selected as its original dataset to be used for creating simulated MAGs. We simulated MAGs from the complete genomes mimicking the distribution of fragmentation, completeness, and contamination observed in UHGG MAGs. Specifically, a list of 100 simulated MAGs was generated using Python SciPy (Virtanen *et al.*, 2020) and NumPy (Van Der Walt *et al.*, 2011) following an F-distribution (see Results). For example, to simulate a 100 MAG dataset with an average fragment number of 50 from 100 complete genomes, we do not cut every genome into 50 fragments. Instead, a Python script is used to generate 100 random numbers with a mean = 50 and following an F-distribution. These 100 random numbers are used to guide the cut of the 100 genomes into fragments to create a simulated MAG dataset. In summary, the number of fragments, and the percentages of completeness and contamination observed in UHGG MAGs were used to randomly generate numbers, which were applied to the 100 complete genomes to create simulated MAG datasets with the expected fragment numbers, completeness percentage, or contamination rates in MAGs. The MAG simulation was performed in three steps (i) fragmentation, (ii) incompleteness, and (iii) contamination (**Figure 1**).

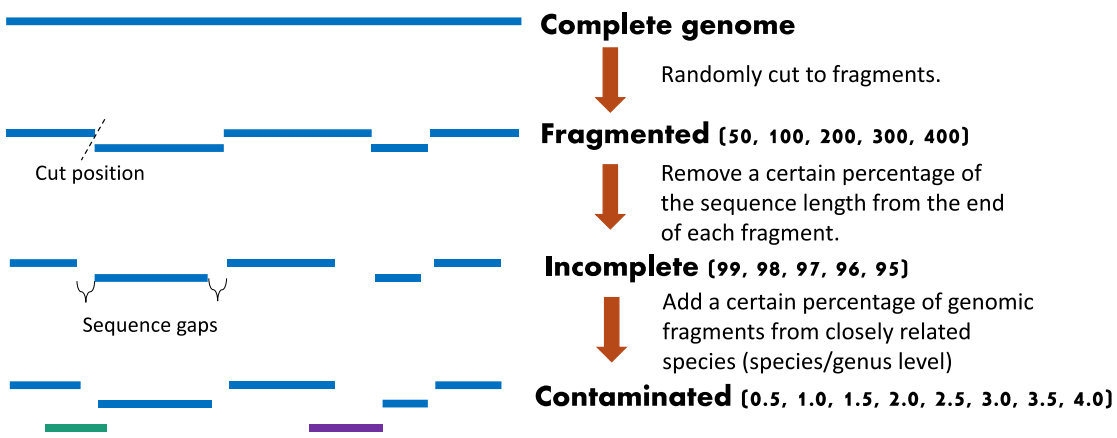

Figure 1. The pipeline to simulate MAGs from complete genomes.

(i) Fragmentation. A total of 5 levels of fragmentation (50, 100, 200, 300, and 400 fragments) were used. Each of the 5 numbers was used as the expected mean value to randomly generate a list containing 100 numbers, corresponding to the number of fragments created in the 100 original complete genomes. More specifically, the numbers of fragments were randomly assigned to complete genomes. Each complete genome was cut according to its assigned number of fragments. The cut positions were randomly selected on the complete genome sequence (number of cut positions = number of fragments - 1). When there were plasmids in the genome, both chromosome and plasmid sequences were cut to generate the fragmented genome. The fragmentation dataset contains 100 fragmented genomes generated as described above.

(ii) Completeness. There were also five levels (99, 98, 97, 96, and 95%) of completeness to be simulated. Each of these five percentages was used as the mean to generate the list containing 100 percentages corresponding to the percentages of completeness created in the 100 original complete genomes. The completeness percentages were randomly assigned to fragmented genomes. The percentage of the

genome to be removed is 100% minus the assigned completeness percentage. The length to be removed on each fragment was calculated from multiplying the removal percentage by the length of the fragment. The removed sequence was discarded from the end of each sequence fragment. The incompleteness dataset containing fragmented and incomplete genomes was generated in this step.

(iii) Contamination. Contamination within species-level (intraspecies) was simulated for all the 17 species, while the contamination within genus-level (interspecies) was only performed for four selected species (*B. pseudomallei*, *B. subtilis*, *K. pneumoniae*, and *S. pyogenes*). The species-level contamination genome fragments were selected from all complete genomes within the species (source genomes are from the same species), whereas the genus-level contamination genome fragments were chosen from other species within the genus (source genomes are from the same genus, see details in Data Collection). For each contamination level, there were 8 levels of contamination percentage, ranging from 0.5% to 4.0%, with a gap of 0.5%. These 8 contamination percentages were used to generate the list of percentages corresponding to the percentages of contamination created in the 100 original complete genomes. The contamination percentages were randomly assigned to fragmented and incomplete genomes generated above. The total contamination length was calculated from multiplying the contamination percentage by the total length of the genome. For each genome, the total contamination length was randomly divided into smaller lengths, the total number of which was varying between 1 and 20. Each smaller length was used as contamination sequence length, and a sequence in that length was copied from randomly selected contamination source genomes in genus-level or species-level and added into the

simulated genome. The contamination dataset containing fragmented, incomplete and contaminated genomes was generated in this step.

Overall, four types of datasets were generated: 1) original complete genome datasets, 2) fragmentation datasets, 3) incompleteness datasets, and 4) contamination datasets. For each species, after the whole simulation process, an original complete genome dataset, 5 fragmentation datasets, 5 incompleteness datasets, and 8 contamination datasets were created.

Pan-genome Analyses for Simulation Datasets

For all the 17 species, pan-genome analyses were performed and compared among species for each type of simulation dataset to determine the effects on core genomes caused by genome fragmentation, incompleteness, and contamination. Genomes in simulation datasets were predicted and annotated using Prokka v1.13 (Seemann, 2014) with default parameters. Pan-genome analyses were carried out using Roary v3.13 (Page *et al.*, 2015) with the parameters '-i 90' (minimum amino acid identity of 90% for a positive match in blastp), '-cd 100' (a core gene defined as 100% presence), '-s' (do not split paralogs), and '-e -n' (create fast core gene alignment with MAFFT (Katoh and Standley, 2013)).

To project the correlation between the number of core genes and the number of fragments or the percentage of incompleteness, the exponential model ($y = e^{(ax+b)}$) was used, where y is the number of the core gene families in pan-genome and x represents the number of fragments or the percentage of completeness. The predicted fitting curves

were plotted using R ggplot2 (Wickham, 2016), and the adjusted-R^2 and P values were calculated.

Random Group Variation Test

Selection of 100 original complete genomes to generate simulated MAGs may affect the analysis results. To assess the variation that may be caused by different random 100 genome datasets, 50 random datasets (each with 100 genomes) were generated going through the same simulation process as described above for *Escherichia coli*. For *Bordetella pertussis, Staphylococcus aureus,* and *Klebsiella pneumoniae*, 30 random datasets (each with 100 genomes) were generated. Therefore, for each of the four species, there will be 50 or 30 of: 1) original complete genome datasets, 2) fragmentation datasets, 3) incompleteness datasets, and 4) contamination datasets. These 50 or 30 datasets of the same type (e.g., 100cut fragmentation group) will be analyzed together. The median, mean, and standard deviation for the number of core genes of the 50 or 30 datasets were calculated. The violin plots were created by using R ggplot2 (Wickham, 2016) to visualize the variations among datasets.

Compare Three Pan-genome Tools

In this study, three pan-genome computational tools were used based on literature search results. Roary was selected due to its high citation and processing ability for large-scale datasets. BPGA v1.3 (Bacterial Pan Genome Analysis tool) (Chaudhari *et al.*, 2016) was selected to represents user-friendly tools, while Anvi'o (Eren *et al.*, 2015) was increasingly used in recent years and designed for analysis and visualization of omics

data. The pan-genome results given by these three tools were compared. To remove the differences that may be caused by gene prediction tools (built in each of the three pan-genome analysis tools), the gene annotation files provided by Prokka v1.13 (Seemann, 2014) were used as inputs for all the three tools. The amino acid sequences in fasta format were used for BPGA, while the gene annotation files in genbank format were used in Anvi'o.

Two representative species, *E. coli* and *B. pertussis*, were selected for pan-genome tool comparison due to their different γ values (indicate more open or close pan-genome). Roary was run with the parameters "-i 90 -cd 100 -s -e -n". In BPGA pan-genome analysis, USEARCH (Edgar, 2010) was selected as the gene clustering tool using sequence identity cut-off 90%, the core genes were defined as genes that present in all the genomes (100%). In Anvi'o pan-genome analysis, "anvi-script-reformat-fasta" was first run to remove the very short contigs (length < 4bp) and unify the contig names for each genome in the dataset. A python script (https://github.com/elizabethmcd/genomes-MAGs-database/blob/master/scripts/genbank-parser.py) was modified to generate a tab-delimited file to define external gene calls from the genbank file from Prokka. The "anvi-gen-contigs-database" was used to generate the contig database for each genome by using external gene calls file. The "anvi-run-ncbi-cogs" annotated genes by searching gene sequences in the contig database against the Clusters of Orthologous Genes (COG) database (Galperin *et al.*, 2021) with Diamond (Buchfink *et al.*, 2015). The "anvi-run-hmms" stored hidden Markov model (HMM) hits in the contig database. The "anvi-gen-genomes-storage" was used to create the genome storage for pangenome analysis. The "anvi-pan-genome" created the pan-genome using Markov Cluster Algorithm (MCL)

(Van Dongen and Abreu-Goodger, 2012) with parameters "--mcl-inflation 10 --use-ncbi-blast --minbit 0.8" (mcl-inflation defines the sensitivity of MCL algorithm during the identification of the gene clusters, minbit defines the minimum bit score ratio between the two amino acid sequences). The "anvi-display-pan" was used to visualize the pan-genome results in the anvi-interactive interface; the state of pan-genome display was saved as "default" and the bins containing all core gene clusters (genes present in all the genomes in the dataset) was saved as "core" collection. The "anvi-summarize" was used to generate the HTML output for pan-genome results.

For the *E. coli* simulation datasets, the Anvi'o pan-genome analysis was also performed by using the default gene prediction tool prodigal (Hyatt *et al.*, 2010) instead of the external gene calls extracted from Prokka (Seemann, 2014) to compare the effects on core genome sizes caused by different gene prediction tools.

Different Parameters in Pan-genome Analysis

Two very important parameters in the pan-genome analysis were investigated: (i) the sequence identity for clustering homologs (e.g., two genes have to be > 90% to be clustered into the same homologous gene cluster) and (ii) the percentage of genomes to be found for defining core genes (e.g., core genes have to be found in 100% of genomes). The simulation datasets for *E. coli* and *B. pertussis* were used to evaluate the effects of different parameter selection on pan-genome analysis results.

Different core gene thresholds were compared in the *E. coli* simulation datasets when using Roary, BPGA, and Anvi'o (use the default prodigal) for pan-genome analysis. The threshold was set as 100%, 99%, 98%, 95%, 92%, and 90% for

fragmentation and contamination dataset groups, and an additional two (88% and 85%) were also used for incompleteness dataset groups. For *B. pertussis,* only fragmentation and incompleteness datasets were used and only Roary was used for pan-genome analysis with different core gene thresholds.

The sequence identity threshold for clustering was tested in the *E. coli* simulation datasets when using Roary. The "-i" parameter was set as 95%, 90%, 85%, and 80%, while the core gene threshold "-cd" was set as 100%, 99%, 95%, and 90%.

Downstream Analysis

Two important downstream analyses are often performed based on the pan-genome results: (i) Clusters of Orthologous Genes (COG) functional analysis and (ii) phylogenetic analysis. The simulation datasets for *E. coli* and *B. pertussis* were used to test the effects on these two types of analysis.

(i) COG analysis.

The representative core/unique gene sequences in the pan-genome of each dataset were extracted and used as queries to search against the COG Conserved Domain Database (CDD) (Lu *et al.*, 2020) using Reversed Position Specific Blast (RPS-Blast) with option '-evalue 1E-5'. Genes having multiple non-overlapping domains were assigned to different COG functional categories and were counted multiple times. The number of core genes assigned to each COG functional category was calculated.

(ii) Phylogenetic tree comparison.

(a). The trees reconstructed using the presence and absence of accessory genes were provided in Roary outputs (accessory_binary_genes.fa.newick). The roary_plots.py

script (https://github.com/sanger-pathogens/Roary/tree/master/contrib/roary_plots) was used to show the presence and absence matrix against a tree.

(b). The core gene nucleotide sequence alignment file (core_gene_alignment.aln) provided in Roary outputs was used to construct the phylogenetic tree based on core genes by using Fasttree v2.1 (Price *et al.*, 2010) with "JCCAT" substitution model (default).

To compare two trees and quantify the differences between them, the normalized Robinson-Foulds symmetric distance (nRF) (Robinson and Foulds, 1981) is a popular metric. The nRFs between the tree built from the original complete genomes and the tree from simulated genomes were calculated by using ETE3 toolkit (Huerta-Cepas *et al*, 2016).

CHAPTER IV

RESULTS

Literature Search and Summary: MAGs are increasingly used in the pan-genome analysis in various ecological environments

The literature search in PubMed and Google Scholar found 18 and 404 papers using MAGs in pan-genome analysis (see Methods), respectively. After manual curation, 40 papers met the filtering criteria. In the recent five years, there was an increasing trend in using MAGs for pan-genome analysis in various research areas, especially in exploring the microbiomes of human, ocean, and hydrothermal vent (**Figure 2**). The pan-genome computational tools including Anvi'o (Eren *et al.*, 2015), Roary (Page *et al.*, 2015), BPGA (Chaudhari *et al.*, 2016), GET_HOMOLOGUES (Contreras-Moreira and Vinuesa, 2013), and OrthoMCL (Li *et al.*, 2003) were used in these 40 studies (**Appendix A**).

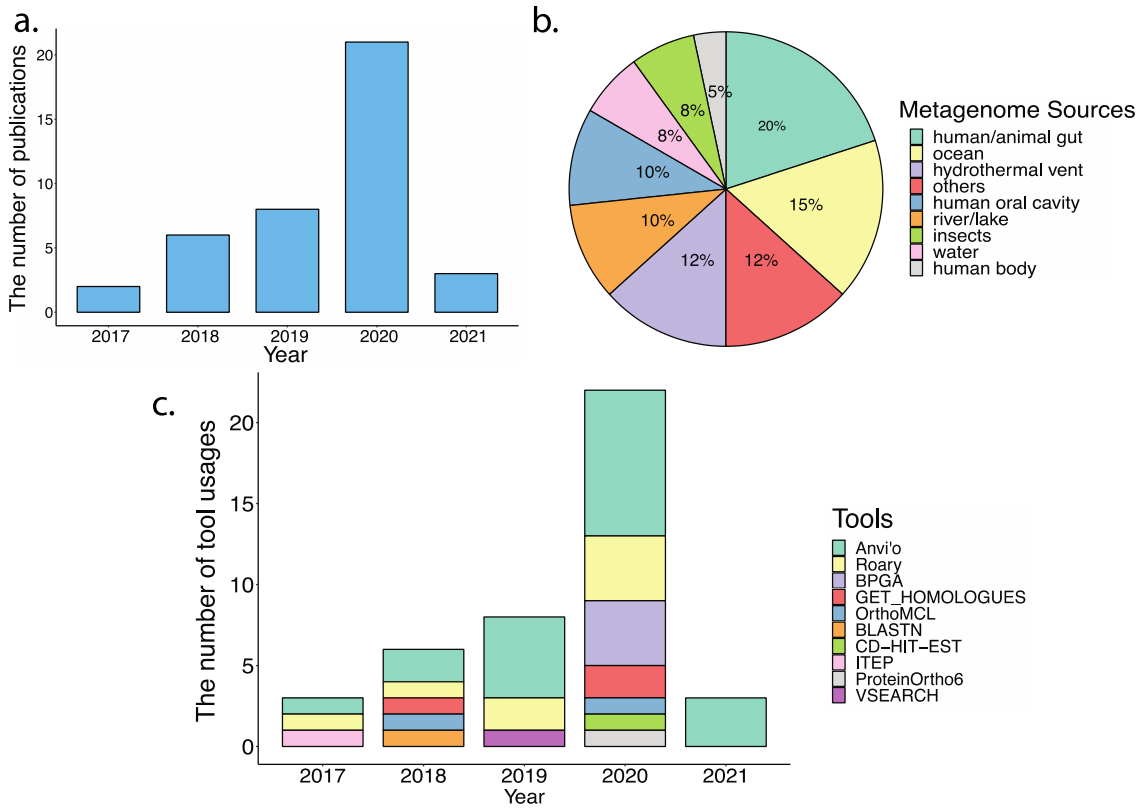

Figure 2. Summary of 40 publications that using MAGs for pan-genome analysis. a: the bar graphs for publication years. **b**: the pie chart for metagenomes sources in publications. **c**: the bar graphs for pan-genome computational tool usages. Multiple tools used in a study are separately counted (**c**).

General Characteristics of 17 Species: the 17 species are from 15 families of 3 phyla and each species have more than 100 complete genomes sharing ANI > 95%

A total of 4,795 complete genomes from 17 species were downloaded. The 17 species belong to 15 taxonomic families of three phyla (**Table 1**): Proteobacteria (10 species), Firmicutes (5), and Actinobacteria (2). Three species, *Escherichia coli*, *Salmonella enterica*, and *Bordetella pertussis*, have the largest numbers of complete genomes in NCBI RefSeq database. The largest average genome size (6,175 genes) was observed in *Pseudomonas aeruginosa*, while there was only an average of 1,546 genes detected in *Helicobacter pylori* genomes (**Table 1**).

Table 1. Summary of 17 prokaryotic species. Pan-genome size and core genome size were calculated using Roary. Core genes were counted as the genes that are present in all the complete genomes in a species. The γ values were calculated by using Heaps' law model.

#	Species	Number of Complete Genomes in NCBI RefSeq	Family	Phylum	Genome Size (Average # Genes)	ANI (%)	Pan-genome Size (# Gene Families)	Core Genome Size (# Gene Families)	γ value (Heap's law*)
1	*Escherichia coli*	923	*Enterobacteriaceae*	*Proteobacteria*	4873	97.78	33123	1486	0.3170
2	*Salmonella enterica*	746	*Enterobacteriaceae*	*Proteobacteria*	4577	98.38	27085	1877	0.3182
3	*Bordetella pertussis*	539	*Alcaligenaceae*	*Proteobacteria*	3901	99.98	4115	2960	0.0357
4	*Staphylococcus aureus*	443	*Staphylococcaceae*	*Firmicutes*	2639	98.43	6177	1710	0.1503
5	*Klebsiella pneumoniae*	372	*Enterobacteriaceae*	*Proteobacteria*	5320	99.04	22106	2741	0.2739
6	*Pseudomonas aeruginosa*	196	*Pseudomonadaceae*	*Proteobacteria*	6175	98.90	19637	3734	0.2669
7	*Streptococcus pyogenes*	195	*Streptococcaceae*	*Firmicutes*	1758	98.77	3959	1267	0.1667
8	*Listeria monocytogenes*	180	*Listeriaceae*	*Firmicutes*	2909	96.96	6109	2098	0.1515
9	*Mycobacterium tuberculosis*	176	*Mycobacteriaceae*	*Actinobacteria*	4067	99.92	4449	3429	0.0316
10	*Campylobacter jejuni*	171	*Campylobacteraceae*	*Proteobacteria*	1746	97.81	4027	1204	0.1865
11	*Helicobacter pylori*	166	*Helicobacteraceae*	*Proteobacteria*	1545	95.25	3755	1143	0.2192
12	*Acinetobacter baumannii*	161	*Moraxellaceae*	*Proteobacteria*	3813	98.28	12426	2132	0.2649
13	*Enterococcus faecium*	115	*Enterococcaceae*	*Firmicutes*	2937	98.39	7652	1608	0.2226
14	*Neisseria meningitidis*	106	*Neisseriaceae*	*Proteobacteria*	2042	98.20	3347	1414	0.1285
15	*Bacillus subtilis*	102	*Bacillaceae*	*Firmicutes*	4178	98.44	8987	3063	0.1901
16	*Burkholderia pseudomallei*	101	*Burkholderiaceae*	*Proteobacteria*	5918	99.41	12941	4663	0.2072
17	*Corynebacterium pseudotuberculosis*	100	*Corynebacteriaceae*	*Actinobacteria*	2127	99.35	2449	1887	0.0370

The number of chromosomes and plasmids in species complete genomes were shown in **Table 2**. Except for *Burkholderia pseudomallei,* all species contained only one main chromosome. The average number of plasmids was greater than 1 in four species (green shading in **Table 2**), where some genomes even contained 11 or 12 plasmids. Some species (grey shading in **Table 2**) had fewer numbers of plasmids, while four species (blue shading in **Table 2**) had no plasmid at all.

Table 2. The number of chromosomes and plasmids in species complete genomes. "Chr" represents chromosome, and "Plas" represents plasmid. "Avg", "Max" and "Min" represent the average, maximum and minimum value, representatively.

Species	Avg#Chr	Max#Chr	Min#Chr	Avg#Plas	Max#Plas	Min#Plas
Enterococcus faecium	1	1	1	4.03	11	0
Klebsiella pneumoniae	1	1	1	3.23	10	0
Escherichia coli	1	1	1	1.89	12	0
Acinetobacter baumannii	1	1	1	1.44	8	0
Burkholderia pseudomallei	2	2	2	0.01	1	0
Salmonella enterica	1	1	1	0.91	7	0
Staphylococcus aureus	1	1	1	0.75	4	0
Bacillus subtilis	1	1	1	0.3	3	0
Campylobacter jejuni	1	1	1	0.23	2	0
Listeria monocytogenes	1	1	1	0.23	5	0
Helicobacter pylori	1	1	1	0.18	2	0
Pseudomonas aeruginosa	1	1	1	0.17	3	0
Streptococcus pyogenes	1	1	1	0.03	3	0
Bordetella pertussis	1	1	1	0	0	0
Corynebacterium pseudotuberculosis	1	1	1	0	0	0
Mycobacterium tuberculosis	1	1	1	0	0	0
Neisseria meningitidis	1	1	1	0	0	0

The average ANI value of 17 species was recorded in **Table 1**, and ANI value distribution of genomes in each species was shown in **Figure 3a**. Species like *B. pertussis* *B. pseudomallei*, and *M. tuberculosis* had average ANI values > 99%, while *L. monocytogenes* and *H. pylori* had average ANI values <97%. Although average ANI values >98% were observed in *S. enterica* and *E. faecium*, some genomes in these species only had pairwise ANI values around 95%, indicating less similarity between genomes.

The phylogenetic relationship of the 17 species predicted using maximum likelihood was shown in **Figure 3b**. In most of the clades, the bootstrap values were 100, indicating high confidence in the phylogenetic species clustering. However, low

bootstrap values were observed when determining the relationship among *Listeria monocytogenes*, *Staphylococcus aureus*, and *Bacillus subtilis*.

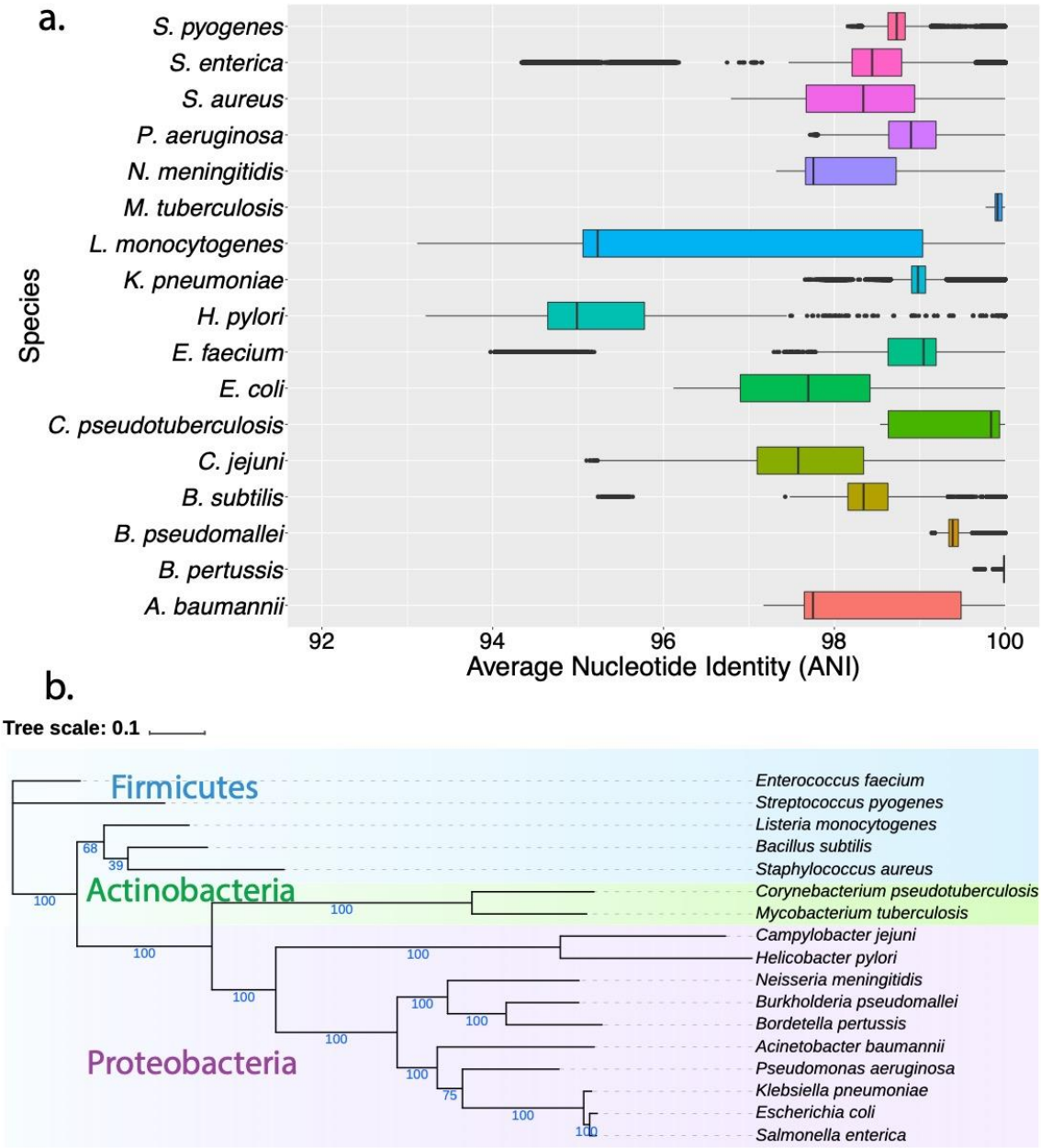

Figure 3. The ANI values and phylogenetic tree for 17 species. a: The boxplots of ANI values for all complete genomes in species. Numbers next to the box represent the average ANI for species. **b:** The phylogenetic tree generated for 17 species representative genomes. Numbers below branches are bootstrap values provided as a percentage over 100 replications.

Species Pan-genome Composition Structure: species with larger number genomes tend to have larger pan-genomes (mostly unique genes)

The pan-genome was built for each species dataset using Roary 3.13.0 with the strictest core gene threshold (genes that were shared by all the genomes in the dataset). All the complete genomes in a species were used as the species dataset. Pan-genome and core genome sizes of species were listed in **Table 1.** The pan-genome size significantly varied among the 17 species. *E. coli, S. enterica* and *K. pneumoniae* all belong to *Enterobacteriaceae* family, and they had large pan-genomes containing > 22,000 gene clusters. In contrast, *C. pseudotuberculosis* and *N. meningitidis* only had <3,400 gene clusters in their pan-genomes. Clearly the number of genomes used in pan-genome construction influence the pan-genome size for a species. There was a positive correlation between the number of genomes and the pan-genome size among different species (Pearson correlation coefficient $R = 0.71$, $p = 0.0013$). There were exceptions when comparing some species. For example, while 101 genomes were used to calculate the pan-genome of *B. pseudomallei*, and 12,941 gene families were identified, only 6,177 gene families were found in 443 *S. aureus* genomes. Moreover, the species pan-genome size was positively correlated with the average genome size ($R = 0.72$, $p = 0.0012$).

The pan-genome composition for each species was shown in **Figure 4**. Obviously, some species tend to have more unique genes, while others have more core genes. A large proportion of cloud genes (genes shared in fewer than 15% genomes in the dataset) were observed in the pan-genome of *E. coli* and *S. enterica*, whereas the pan-genome of *B. pertussis* consisted of >70% (2960/4115) core genes. *Campylobacter jejuni and S. pyogenes* showed similar pan-genome composition, and their pan-genome

sizes were ~4000 gene families which contain ~1000 core gene families. Similar pan-genome sizes were found in *B. pseudomallei* and *A. baumannii*; however, the core gene size in the former was about twice as large than that in the latter.

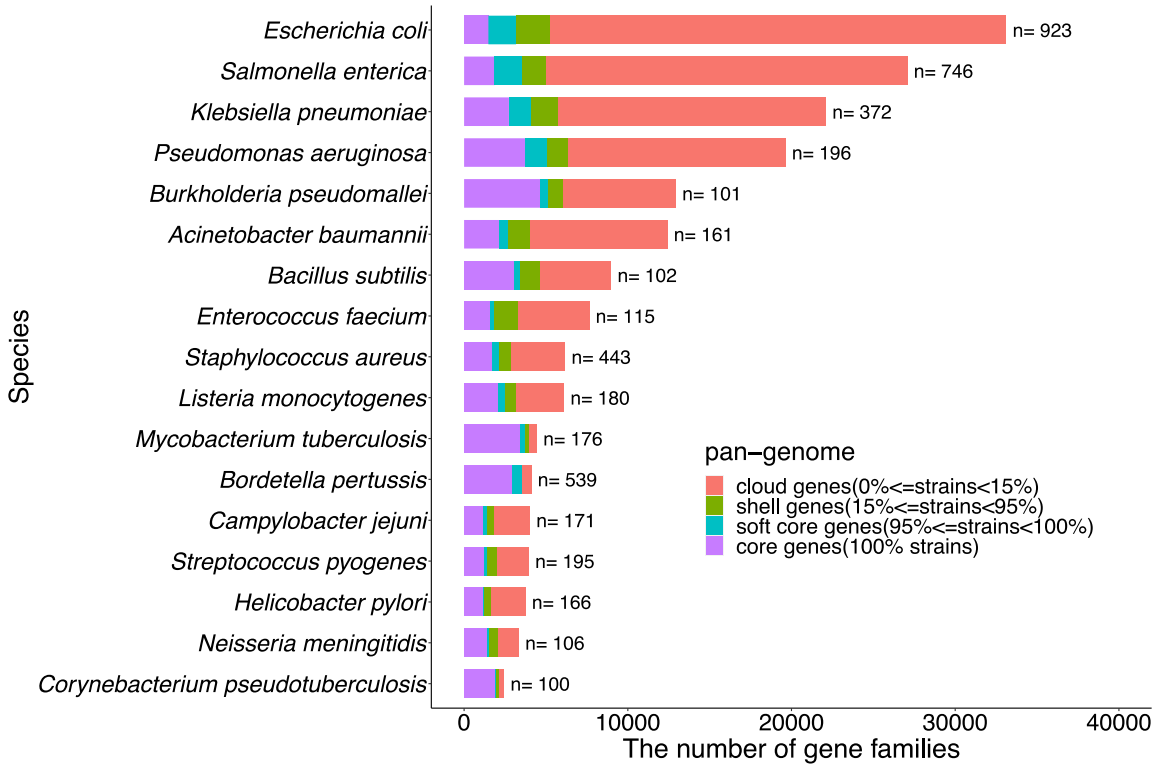

Figure 4. The Pan-genome composition structure of 17 species constructed by using Roary. The n beside each bar represents the number of complete genomes used for building species pan-genome. Core genes that are found in all the genomes are shown in purple. Soft core genes that are present in at least 95% but not all of the genomes are shown in cyan. Shell genes that can be observed in more than 15% but fewer than 95% of the total genomes are shown in green. Cloud genes that are only present in fewer than 15% of the total genomes are shown in red.

The Heaps' law model was used to calculate the exponent γ value for each species to predict the openness and closeness of the pan-genome (**Table 1**). If γ >0, the size of the pan-genome would increase with the addition of newly sequenced genomes, and the pan-genome was identified as open; otherwise, the pan-genome was identified as closed (Tettelin *et al.*, 2008). All 17 species had positive γ values, indicating their open pan-

genomes. The two largest γ values, 0.3182 and 0.3170, were found in the pan-genomes of *S. enterica* and *E. coli*, respectively. On the other hand, three species (*B. pertussis, M. tuberculosis,* and *C. pseudotuberculosis*) only had γ values at ~0.03, which was close to 0. The lower γ values reflected that their pan-genomes were near to be closed.

Statistical Analysis of MAGs: the observed distribution of three metrics is used to guide the creation of simulated MAGs from complete genomes

All MAGs (N=276,349) in Unified Human Gastrointestinal Genome (UHGG) collection (Almeida *et al.*, 2021) were used to determine the distribution of contig fragment number, completeness, and contamination in real MAGs (**Figure 5**). Overall, the number of fragments in 276,349 MAGs varied from 1 to 2282 with a mean = 208. The MAG completeness was distributed between 50% and 100% with a mean = 85.18%, while the contamination varied from 0% to 5% with a mean = 1.2%.

The 276,349 UHGG MAGs were assigned to 3,751 species by EBI, which include 8 of the 17 species we collected from the RefSeq database for complete isolate pan-genome analysis (**Figure 4**). When looking at MAGs of 5 species (each species has more than 50 genomes in UHGG MAGs), they all had different distributions (**Figure 6**), but these were all skewed distributions. Therefore, F distribution (a theoretical distribution in Statistics often used to model skewed distributions) was applied to generate random number sets to simulate the distribution of fragment numbers, completeness, and contamination by adjusting the parameters d1 and d2.

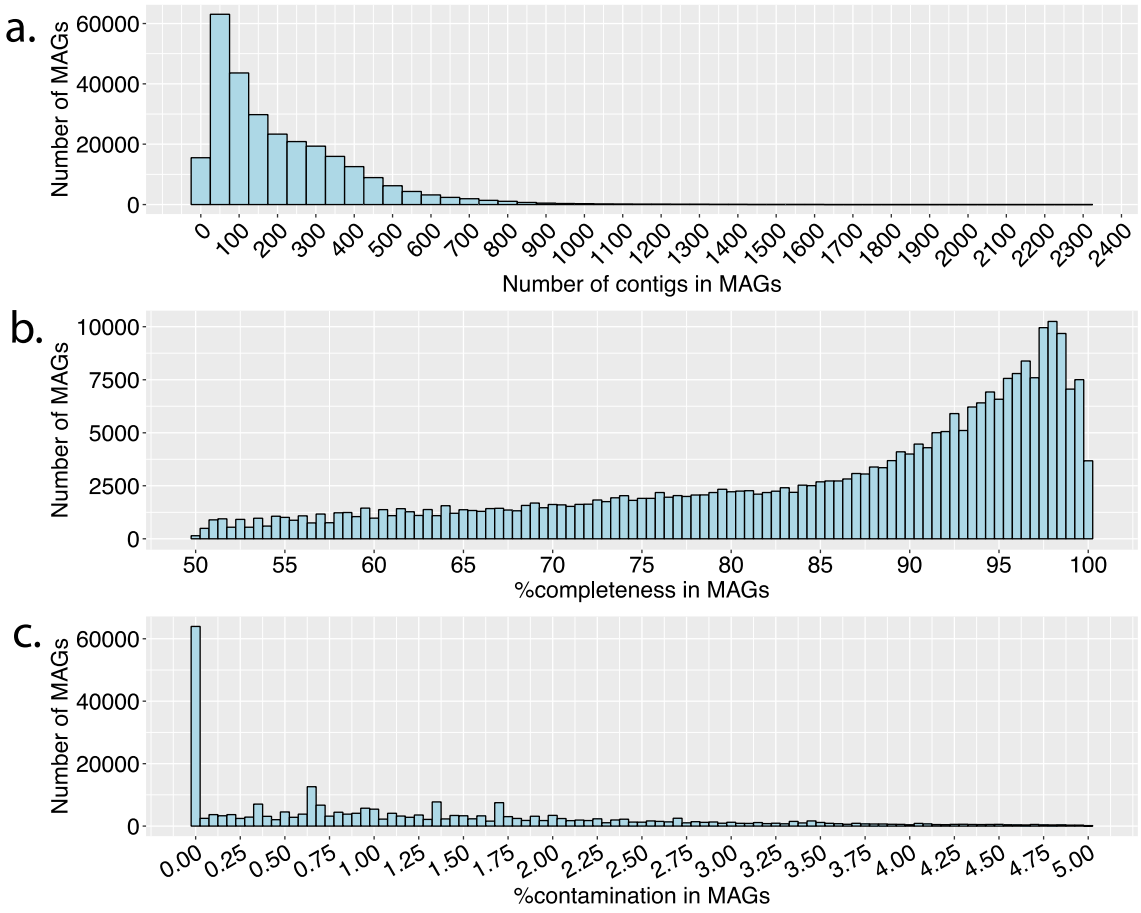

Figure 5. Histogram of MAG statistical analysis. The distribution of contig numbers (**a**), completeness percentage (**b**), and contamination rates (**c**) in 276,349 MAGs. All the MAGs are extracted from UHGG built by Almeida et al. Histogram a and c show a skewed distribution to the right, while histogram b shows a skewed distribution to the left.

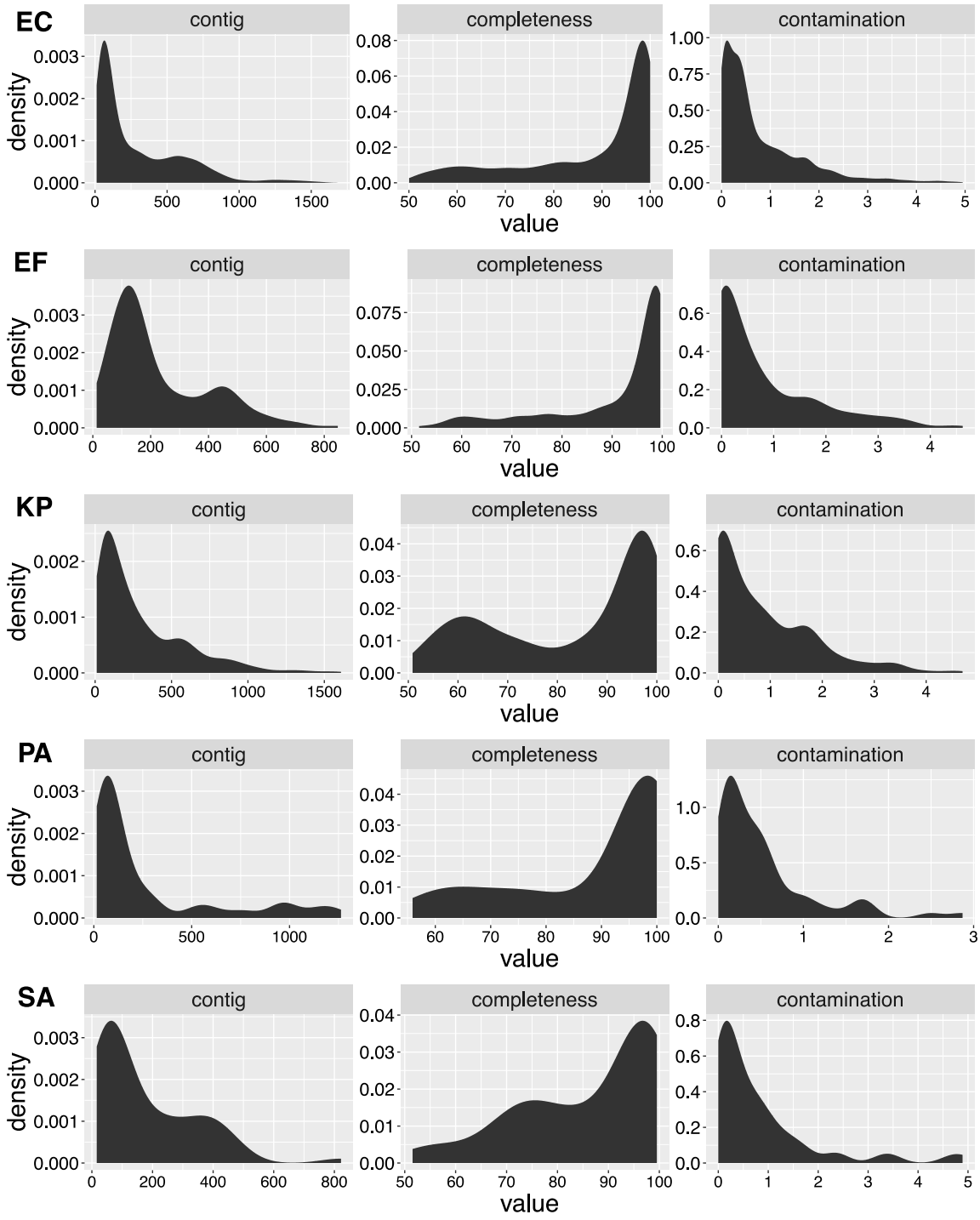

Figure 6. The density plots of UHGG MAG statistical analysis for five species. EC represents *Escherichia coli* (n=4391), EF represents *Enterococcus faecium* (n=333), KP represents *Klebsiella pneumoniae* (n=641), PA represents *Pseudomonas aeruginosa* (n=63), and SA represents *Staphylococcus aureus* (n=57).

More Fragmented genomes Led to More Core Gene Loss

We have created simulated MAGs from the RefSeq complete genomes following the F distribution of the three metrics (see Methodology). For each of the 17 selected species (**Table 1**), 100 complete genomes were randomly selected as the original genome dataset, and each genome is randomly fragmented. If we intend to have the 100 complete genomes cut into an average of 50 fragments, 50 is the mean value, which is used to generate a set of 100 numbers with a mean = 50 following an F distribution. The 100 numbers are randomly assigned to the 100 original genomes as the number of fragments that need to be generated for the resulting fragmented genomes (i.e., simulated MAGs). and then used to make the random cuts in the 100 original genomes. The original 100 genomes and the simulated MAGs were subjected to pan-genome analyses separately, and the results were compared to evaluate the effect of fragmentation on pan-genome analysis.

E cherichia coli has an open pan-genome, while *B. pertussis* has closed pan-genome. Therefore, they were selected as representative species to show the comparison results between original genomes and fragmented genomes. **Figure 7a** shows that the number of the core gene families significantly decreased with the increasing number of fragments in *E. coli*. There were more than 2,600 core gene families in the original dataset; however, only 491 gene families were shared by all the genomes when the 100 *E. coli* genomes had an average of 400 fragments. Similar core gene loss results were observed in *B. pertussis*. As shown in **Figure 7b**, more than 75% of core gene families in *B. pertussis* original genomes were lost due to 400 average fragments in genomes.

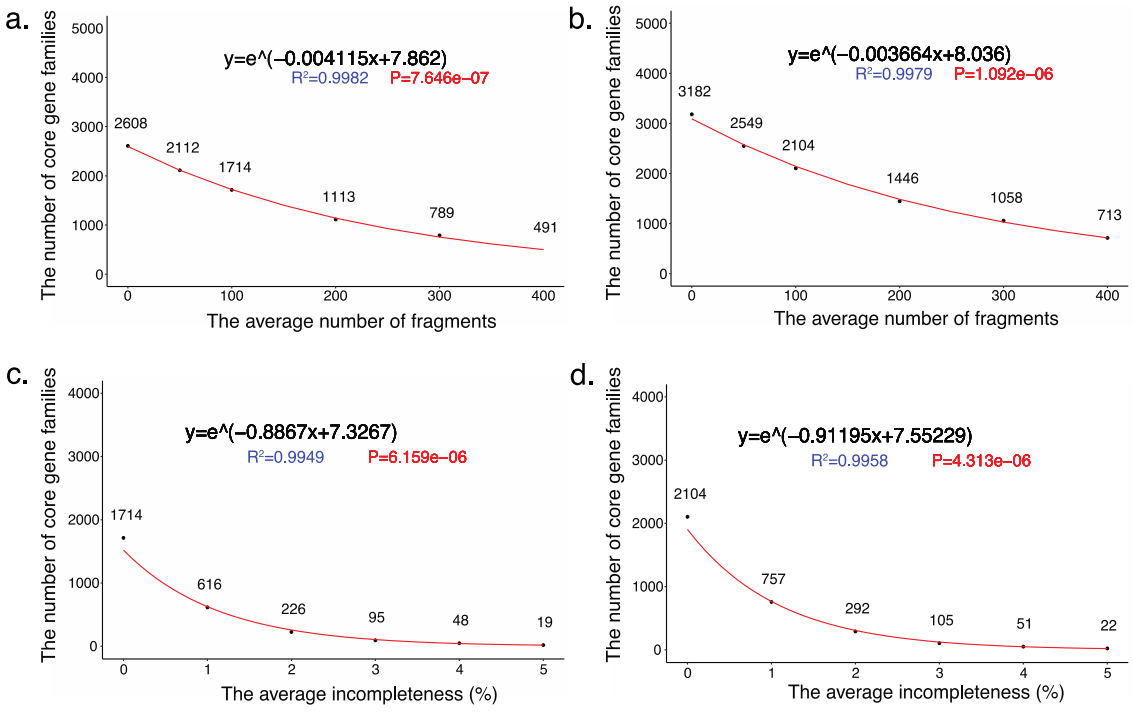

Figure 7. Core genome sizes decrease in *Escherichia coli* and *Bordetella pertussis* genomes. a&b: The curve of core genome sizes in 100 *E. coli* genomes (**a**) and 100 *B. pertussis* genomes (**b**) predicted as a function of the average number of fragments in species genomes. **c&d:** The curve of core genome sizes in 100 *E. coli* genomes (**c**) and 100 *B. pertussis* genomes (**d**) predicted as a function of the average incompleteness in species genomes. Black dots represent the actual number of core gene families in 100 genomes with the varying average number of fragments or average incompleteness. Red curves are predicted by using the exponential model. The equation for the species exponential curve is shown in black font. R^2 and P-value of the predicted curve are shown in blue and red font, respectively.

Given fragmentation, the reduction of core genome sizes was also observed in other 15 species (**Figure 8**). For instance, the number of the core gene families dropped from ~4600 to <1900 in *B. pseudomallei*. A greater number of core gene families were lost during genome fragmentation in species having larger core genome size in the original genomes. In comparison, species like *E. faecium* had a smaller size core genome before fragmentation, the number of core gene families was <500 after an average of 400 fragments in genomes. Varying loss of core genome was observed in different species.

The exponential model was used to quantitatively evaluate the degree of core genome loss in species (**Figure 7 a, b and Figure 8**):

$$y = e^{(ax+b)},$$

where y was the number of the core gene families in the pan-genome and x represented the average number of fragments in genomes used to build the pan-genome. The parameter a and b were used to show how the core genome size would change with the average number of fragments in genomes.

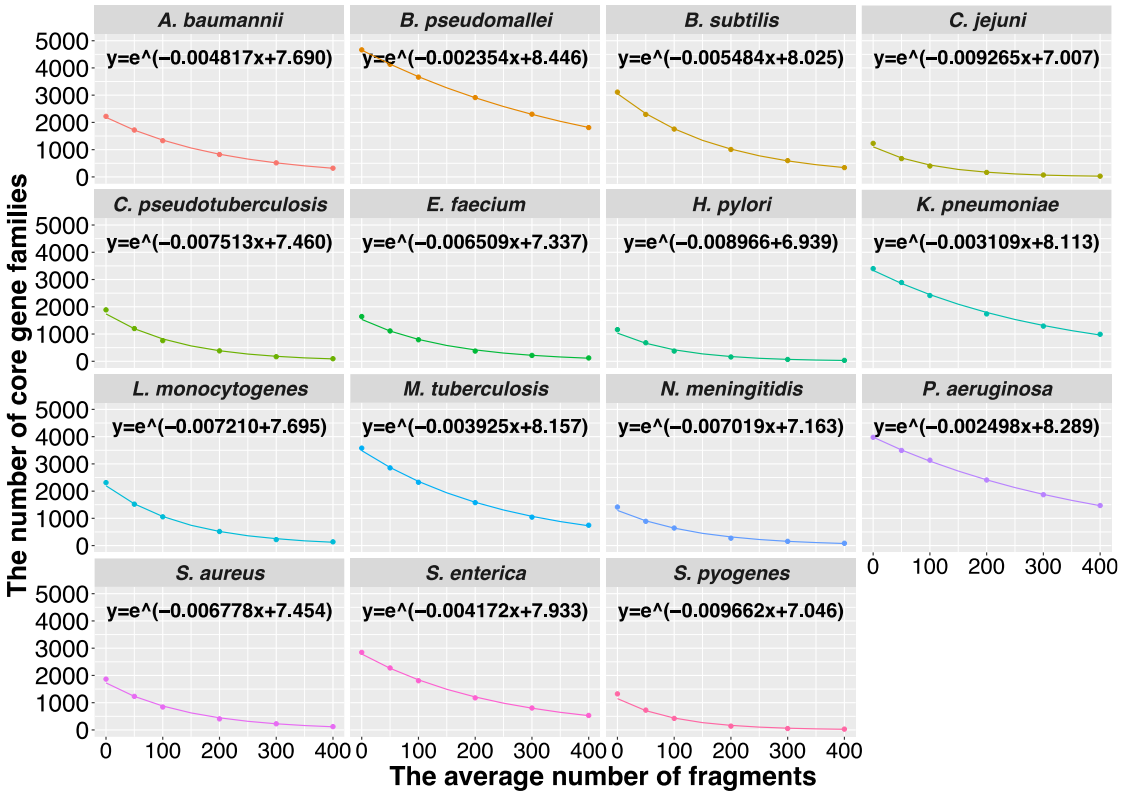

Figure 8. Fragmentation effects on the number of core gene families. The curve of the number of core gene families in 100 genomes of each species predicted as a function of the average number of fragments in species genomes. The dots in each species plot represent the actual number of core gene families in 100 genomes with the varying average number of fragments. The curves are predicted by using the exponential model. The equation for the species curve is shown in black font.

The fitted curves of all species have an adjusted R^2 >0.99 with P-values < 0.0001 (**Table 3A**). The fitted curves could be applied well to predict the number of core genes under a specific average number of fragments in 100 genomes. Species like *B. pseudomallei* and *P. aeruginosa* had parameter |a| < 0.0025 and parameter b > 8.2. In contrast, the parameter |a| > 0.009 and parameter b ~ 7.0 were found in *H. pylori* and *C. jejuni*. In general, species with a large number of core gene families in the original genomes have smaller |a| and larger b.

Table 3. The exponential equation table for 17 species. Adjusted R^2 and P-values for each species in fragmentation (A) and incompleteness (B) simulation.

A

Species	Equation	Adjusted.R.squared	P.value
B. pseudomallei	y=e^(−0.002354x+8.446)	0.9999	1.44e−09
B. subtilis	y=e^(−0.005484x+8.025)	0.9997	2.02e−08
A. baumannii	y=e^(−0.004817x+7.690)	0.9996	3.19e−08
P. aeruginosa	y=e^(−0.002498x+8.289)	0.9996	3.73e−08
S. enterica	y=e^(−0.004172x+7.933)	0.9989	2.73e−07
M. tuberculosis	y=e^(−0.003925x+8.157)	0.9983	7.15e−07
E. coli	y=e^(−0.004115x+7.862)	0.9982	7.65e−07
B. pertussis	y=e^(−0.003664x+8.036)	0.9979	1.09e−06
C. jejuni	y=e^(−0.009265x+7.007)	0.9971	2.01e−06
K. pneumoniae	y=e^(−0.003109x+8.113)	0.9970	2.21e−06
S. aureus	y=e^(−0.006778x+7.454)	0.9954	5.19e−06
C. pseudotuberculosis	y=e^(−0.007513x+7.460)	0.9948	6.48e−06
E. faecium	y=e^(−0.006509x+7.337)	0.9946	6.92e−06
H. pylori	y=e^(−0.008966+6.939)	0.9933	1.09e−05
L. monocytogenes	y=e^(−0.007210+7.695)	0.9932	1.12e−05
N. meningitidis	y=e^(−0.007019x+7.163)	0.9917	1.66e−05
S. pyogenes	y=e^(−0.009662x+7.046)	0.9820	7.79e−05

B

Species	Equation	Adjusted.R.squared	P.value
A. baumannii	y=e^(−0.88546x+7.12216)	0.9984	6.44e−07
M. tuberculosis	y=e^(−0.98537x+7.70595)	0.9981	8.53e−07
S. enterica	y=e^(−0.96558x+7.47089)	0.9979	1.09e−06
K. pneumoniae	y=e^(−0.95979x+7.73514)	0.9971	2.08e−06
B. subtilis	y=e^(−0.9925x+7.40796)	0.9968	2.43e−06
P. aeruginosa	y=e^(−0.91345x+8.01820)	0.9967	2.68e−06
B. pertussis	y=e^(−0.91195x+7.55229)	0.9958	4.31e−06
B. pseudomallei	y=e^(−0.90748x+8.12760)	0.9956	4.72e−06
E. coli	y=e^(−0.8867x+7.3267)	0.9949	6.16e−06
N. meningitidis	y=e^(−0.8134x+6.38212)	0.9943	7.78e−06
C. pseudotuberculosis	y=e^(−1.0970x+6.7790)	0.9943	7.87e−06
S. pyogenes	y=e^(−0.9319x+5.9872)	0.9861	4.65e−05
L. monocytogenes	y=e^(−0.9735x+6.9558)	0.9852	5.29e−05
H. pylori	y=e^(−0.73620+5.73546)	0.9818	7.98e−05
C. jejuni	y=e^(−1.00402x+6.02742)	0.9787	1.10e−04
S. aureus	y=e^(−0.99026x+6.70034)	0.9782	1.15e−04
E. faecium	y=e^(−1.0442x+6.7452)	0.9462	7.04e−04

More Incompleteness Genomes Had More Core Gene Loss

The incompleteness simulation was performed based on genome datasets after fragmentation, i.e., applying the fragmentation on the 100 original genomes (on average 100 fragments per genome) and then applying the incompleteness. The number of core gene families significantly changed due to the loss of average completeness in *E. coli* and *B. pertussis* genomes (**Figure 7 c, d**). Only 36% (616/1,714 in *E. coli* and 757/2,104 in *B.*

pertussis) core gene families were retained in these two species when an average of 1% genome sequences was removed from each of the genomes. Similar results were observed in all other species regardless of their original core gene sizes (**Figure 9**). When average incompleteness in genomes was 1%, the core genome size in *B. pseudomallei* dramatically changed from >3,600 to <1,500. Likewise, more than 500 gene families in ~800 core gene families in *S. aureus* were missed. Overall, only 1% loss in average completeness, from 100% to 99%, would lead to > 60% loss of core gene families in most species. Species had larger core genomes were more affected than those with smaller core genomes in the original genomes.

All the species would have their core genome size smaller than 50 and even near 0 when the average incompleteness reaching 5% (**Figure 7 c, d, and Figure 9**). That means only a 5% loss in genome completeness would almost lead to the loss of all core genes. Compared to the influence from fragmentation, the incompleteness had more significant effects on the core genome size.

The exponential model was applied to perform curve fitting to show the relationship between the number of core gene families and the average incompleteness in genomes (**Figure 7 c, d, and Figure 9**). Here, x was the average incompleteness in genomes. Most of the fitted curves had adjusted R^2 >0.98 and p-values < 0.0001 (**Table 3B**), however, the curve of *E. faecium* had a lower adjusted R^2 (~0.94) than other species, indicating the curve was less fitted. The fitted curves may be applied well to predict the number of core genes under specific average incompleteness in genomes.

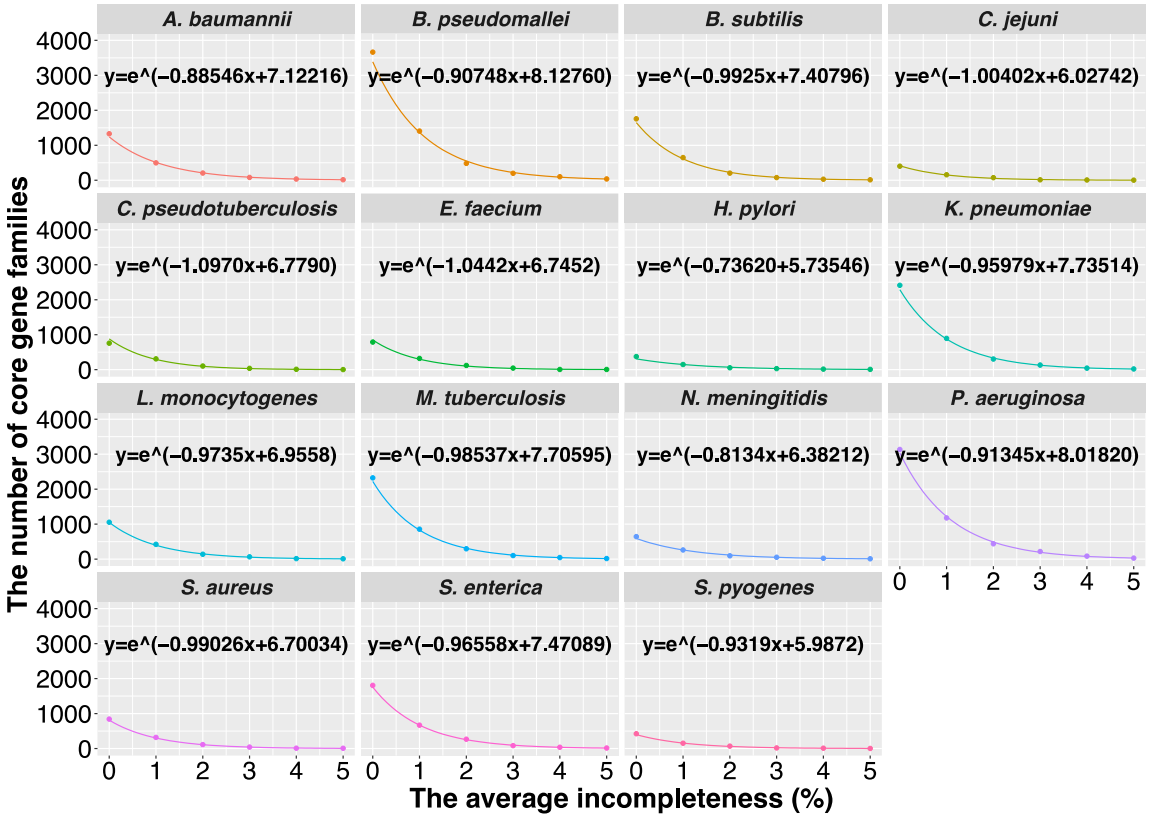

Figure 9. Incompleteness effects on the number of core gene families. The curve of the number of core gene families in 100 genomes of each species predicted as a function of the average incompleteness (%) in species genomes. The dots in each species plot represent the actual number of core gene families in 100 genomes with varying average incompleteness. The curves are predicted by using the exponential model. The equation for the species curve is shown in black font.

Contamination is not supposed to lead to Core Gene loss but in Roary result it is

The contamination simulation was performed based on genome datasets after fragmentation and incompleteness, *i.e.*, applying the contamination on the 100 fragmented and incomplete genomes (on average 100 fragments per genome, and average 1% incompleteness) and then applying the contamination. In binning metagenome contigs based on sequence compositions to form MAGs, the contamination is most likely from closely related genomes (e.g., strains of the same species or genus) as they are more likely to have very similar sequence compositions. Therefore, intra- and inter-species

contamination were added to fragmented and incomplete genomes (See methodology about contamination). The number of core gene families slightly changed with increasing average intraspecies contamination in genomes (**Figure 10**). Unlike the large changes caused by fragmentation and completeness, only seven species (*B. pertussis, B. pseudomallei, E. coli, K. pneumoniae, M. tuberculosis, P. aeruginosa,* and *S. enterica*) were influenced by contaminants with a small change of core genome size. About 500 core gene families were lost in each of the seven species due to an average of 4% contamination. However, in other species, only a slight decrease was observed when the average contamination increased from 0.5% to 4%. This slight decrease was not even noticeable in four species including *C. jejuni, H. pylori, N. meningitidis,* and *S. pyogenes*.

Intuitively, unlike fragmentation and incompleteness, contamination will add additional genes into the pan-genome. If the core gene families decreased, other types of genes will increase. Indeed, for most species, the number of cloud genes (a term used in Roary: genes that are shared by <15% genomes in the dataset) increased constantly when more and more contamination was introduced (**Figure 10**). The most dramatic change was observed in *B. pseudomallei*, whose core genes were also most influenced by contamination. There was about an average of 250 gene additions in the number of cloud gene families for each 0.5% contamination in *B. pseudomallei*. However, the cloud gene families in the four species that have almost constant core genome sizes were less influenced by the contamination. The near-horizontal lines with few fluctuations were seen in these four species (**Figure 10**), indicating the negligible changes in the cloud genome sizes. Clearly, the drop of core genome size due to contamination is correlated with the rise of the variable genome size.

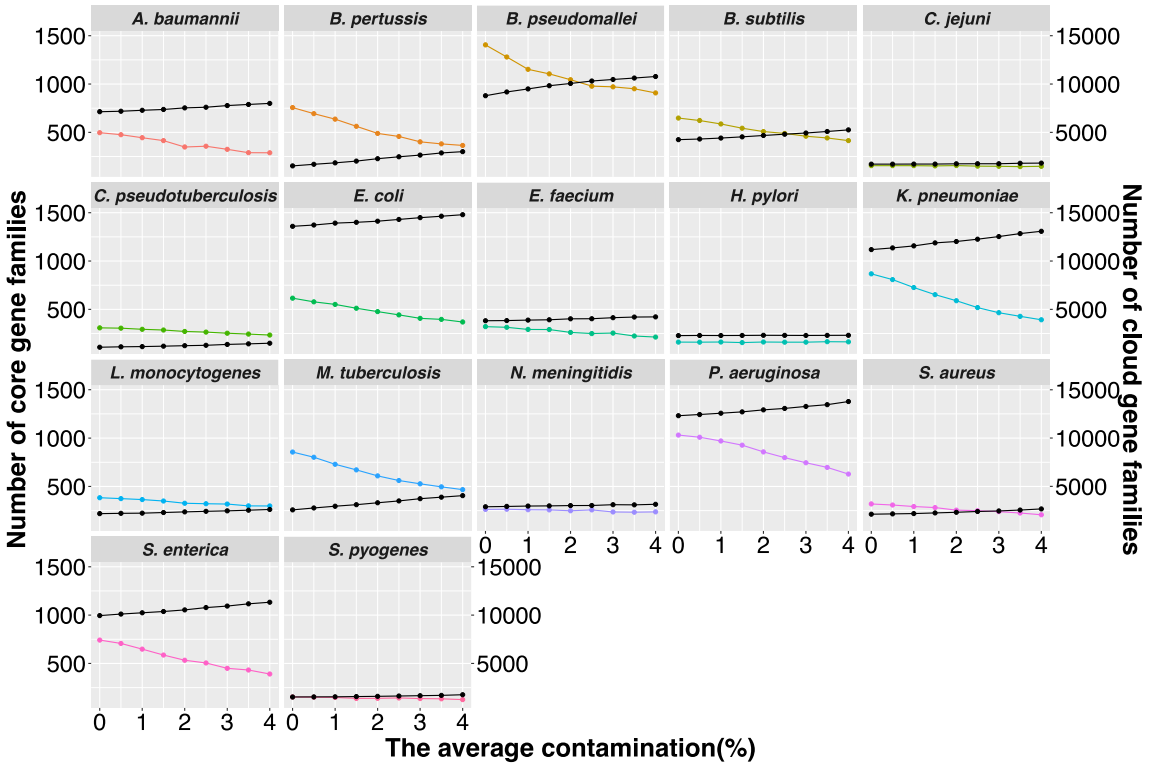

Figure 10. Intraspecies contamination effects on the number of core and cloud gene families. The number of core gene families (colorful, left y-axis) and cloud gene families (black, right y-axis) in 100 genomes of each species changed with average intraspecies contamination (%) added in species genomes.

In addition to intraspecies contamination, to determine the effect of interspecies contamination on the number of core and cloud gene families, four species (*B. pseudomallei*, *B. subtilis*, *K. pneumoniae*, and *S. pyogenes*) were selected as representative species based on their result in previous intraspecies contamination shown in **Figure 10**. Genomes from different species but within the same genus were collected as interspecies contamination (see details in methodology). Unlike intraspecies contamination, when interspecies contamination was introduced, the number of core gene families changed only slightly, whereas the number of cloud gene families increased significantly (**Figure 11**). The cloud genome sizes in *B. pseudomallei* and *B. subtilis*

dramatically increased when the average interspecies contamination changed from 0% to 4%, however, the core genome size was almost constant. Fewer differences in the size of core and cloud genomes caused by intra- and inter-species contamination were observed in *K. pneumoniae*. Although core genome size in *S. pyogenes* was almost the same when adding intra- and inter-species contaminants, the cloud genome size increased with more interspecies contaminants. Since a fewer reduction in species core genome sizes was observed when adding interspecies contamination than intraspecies contamination, the core genome size estimation would be slightly affected by genomic sequences from more distantly related species. However, the results suggested that interspecies contamination increased the number of cloud gene clusters in species, which may lead to possible overestimation in pan-genome size. The smaller decrease in core genes but higher increase in cloud genes from inter-species than intra-species contamination is likely because the lower sequence similarity from inter-species homologs is more likely to put newly added genes from different species to cloud genes.

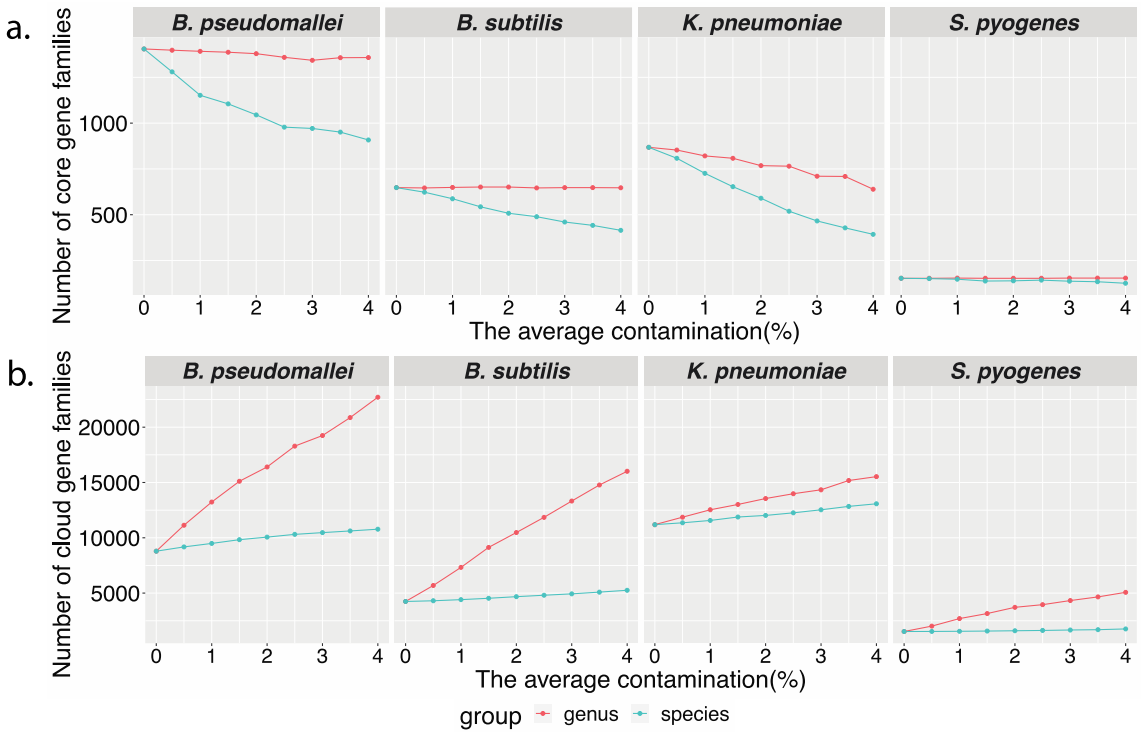

Figure 11. Interspecies and intraspecies contamination comparison. The number of core gene families (**a**) and cloud gene families (**b**) in 100 genomes of four species changed with average intraspecies or interspecies contamination (%) added in species genomes. Group labeled as genus represents the interspecies contamination from other species in the same genus, and group labeled as species represents the intraspecies contamination from different strains in the species dataset.

The core gene loss remains when different sets of random genomes are used

So far, we have only used one dataset of 100 random-selected complete genomes to generate simulated datasets (fragmented, incomplete, and contaminated datasets) for each species. Clearly, choosing which 100 complete genomes to use as the starting point for simulation may affect the simulated datasets and pan-genome analysis results. To assess this effect, we have used 4 species that have more than 200 complete genomes in their corresponding species datasets (*E. coli*, *B. pertussis*, *S. aureus* and *K. pneumoniae*) to generate multiple datasets of 100 random-selected complete genomes for each species,

and repeated the simulations and pan-genome analysis on each of the datasets. **Figure 12a** shows the variations in the number of core gene families among 50 *E. coli* datasets (each with 100 random-selected complete genomes and their simulated genomes) during the simulation process. The median and mean values for core genome size in 100 complete *E. coli* genomes in 50 datasets were 2588 and 2577, with the standard deviation at 112.89. There was an average of >400 core gene family loss due to 50 fragmentation, and another average of >1300 core gene family loss due to 1% loss in genome completeness. On average, about 200 gene reductions in core genome size were noticed after adding 2% contamination. These findings were consistent with the results in one *E. coli* dataset randomly selected from *E. coli* species dataset. During three steps of simulation, the standard deviation among 50 datasets decreased from 112.89 in original datasets to 37.51 in contamination datasets, indicating reduced variations among datasets during simulation. Overall, the decrease of core gene families in simulated genomes is independent of which 100 random genomes were selected.

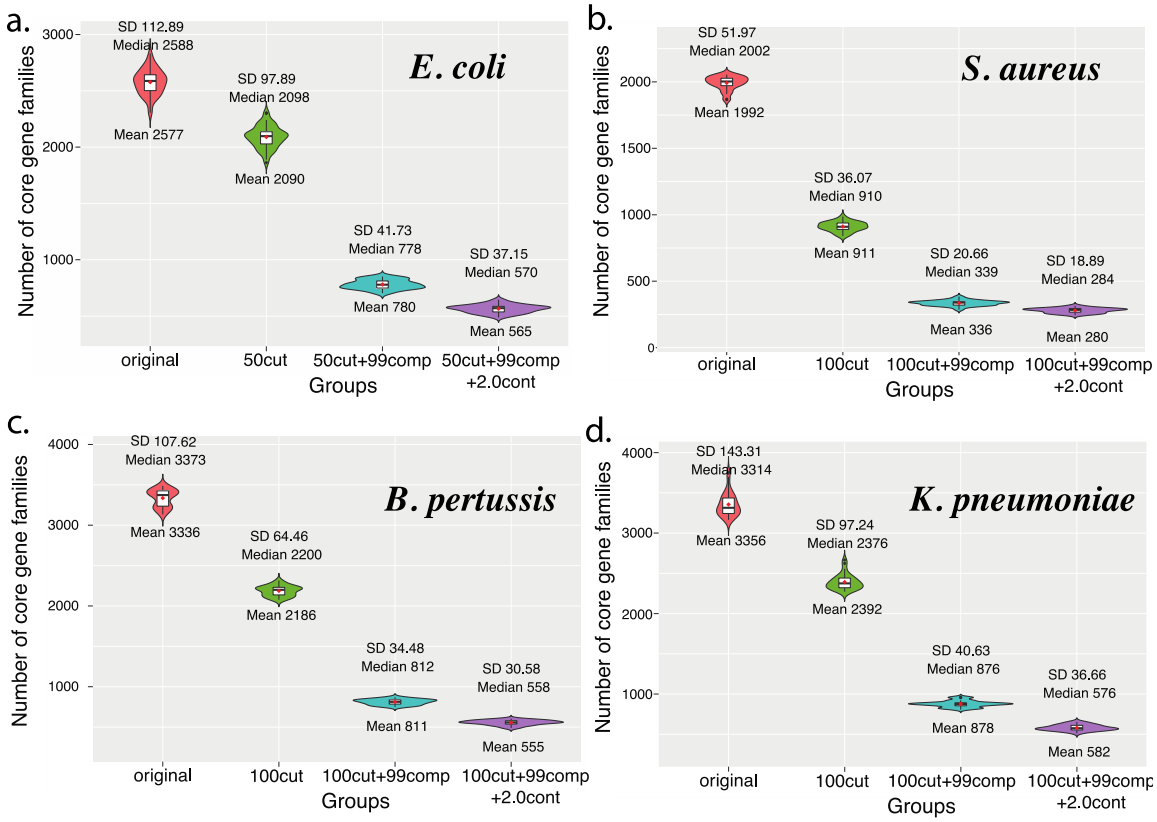

Figure 12. Pan-genome analysis for multiple random simulation datasets. Violin plots of the number of core gene families calculated by Roary in 50 *E. coli* datasets (**a**), 30 *S. aureus* datasets (**b**), 30 *B. pertussis* datasets (**c**), and 30 *K. pneumoniae* datasets (**d**). Groups: "ori" represents the original dataset containing 100 genomes randomly selected from *E. coli* complete genomes (species dataset); "50cut" or "100cut" represents that all genomes in a dataset have an average of 50/100 fragments; "50cut+99comp" or "100cut+99comp" means that genomes in a dataset have an average of 99% completeness based on 50 or 100 fragmentation; "50cut+99comp+2.0cont" or "100cut+99comp+2.0cont" represent that genomes in a dataset have an average of 2.0% intraspecies contamination based on 50 or 100 fragmentation and 99% incompleteness.

Compared to the largest *E. coli* species dataset (923 genomes), smaller species datasets were used for *B. pertussis* (539 genomes), *S. aureus* (443 genomes), and *K. pneumoniae* (372 genomes). Thirty datasets (each with 100 random-selected complete

genomes and their simulated genomes) were used to test the variations among datasets. The core genome size variations within different simulation groups in three species were shown in **Figure 12 b, c, d**. *S. aureus* had the smallest standard deviations among 30 datasets during the simulation process, indicating more conserved core genes in species and few effects caused by random selection (**Figure 12b**). In contrast, *B. pertussis* and *K. pneumoniae* had larger standard deviations in each simulation group, indicating larger variations in genomic contents among species genomes (**Figure 13 c, d**). In *K. pneumoniae*, the group mean value was always higher than the group median value, whereas the reverse results were observed in *B. pertussis*. It remains true that in all these species, the decrease of core gene families in simulated genomes is independent of which 100 random genomes were selected. Therefore, irrespective of what species and what genomes are used, pan-genome analysis using genomes with fragmentation, incompleteness, and contamination (i.e., features of MAGs) will suffer from a significant loss of core gene families.

The core gene loss remains when different pan-genome analysis tools were used

To determine whether the core gene loss was due to the use of Roary for pan-genomic analysis, we have repeated all the analyses using two other popular tools: Bacterial Pan Genome Analysis tool (BPGA) (Chaudhari *et al.*, 2016) and Anvi'o (Eren *et al.*, 2015). Ten datasets were selected from 50 *E. coli* datasets and 30 *B. pertussis* datasets to run BPGA and Anvi'o. The original, fragmentation, incompleteness, and contamination simulation groups were used. The number of core gene families in 10 *E. coli* datasets in each test group were shown in **Figure 13a.** The overall results given by the three tools were consistent in most of the groups except the contamination group

(**Figure 13a**). The core genome loss caused by fragmentation and completeness loss can be observed from results given by three pan-genomic tools. The number of core gene clusters classified by Roary and Anvi'o were a little higher than that identified in BPGA, which may be caused by the different clustering algorithms used in the tools. However, compared to the core genome size reduction given by Roary in contamination groups, the number of core gene families was increased slightly in the results given by BPGA and Anvi'o when an average of 2% contamination was added to genomes. Similar results were observed in *B. pertussis* (**Figure 13b**). Since no core genes were removed from genomes in contamination simulation, the core genome size was not expected to be reduced. The core genome size reduction in the contamination group indicated possible bias in the gene clustering of Roary.

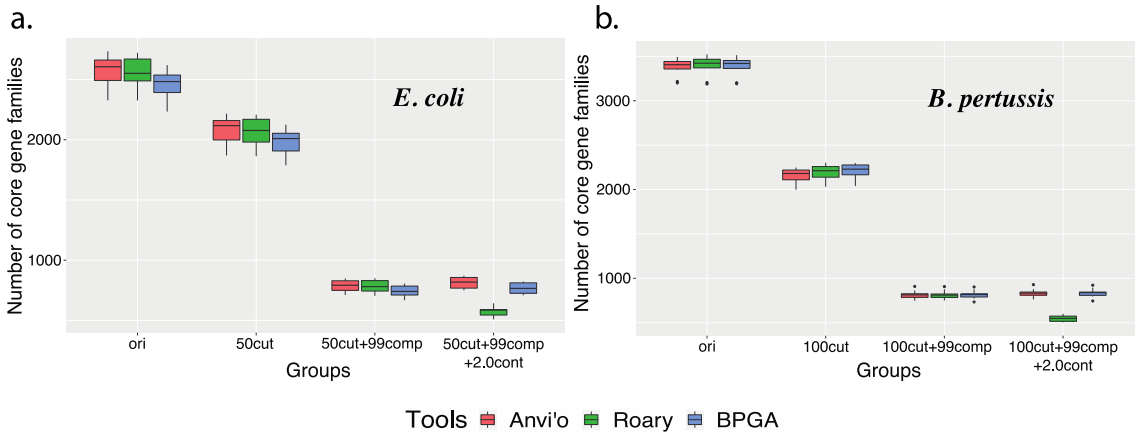

Figure 13. Pan-genome analysis by using different tools. Boxplots of the number of core gene families calculated by Anvi'o (red), Roary (green), and BPGA (blue) in10 *E. coli* datasets (**a**) and *B. pertussis* datasets (**b**). See legend in **Figure 12**.

To further look into the reduction of core genes by Roary before and after 2% contamination was introduced, we have compared the two core gene sets. The result

indicated that some genes identified as core genes before contamination were misclassified into soft core genes (genes present in at least 95% but not all of the genomes) after contamination. In other words, in some genomes their core genes were clustered with other genes with high sequence identity to form a new gene family, leading to the loss of some core gene families (**Appendix B**). As shown below, changing the core gene threshold (i.e., lower the percentage of genomes that need to contain the gene to below 100%) helps address this issue.

The core gene loss can be partially alleviated by lowering the core gene threshold used in pan-genome analysis

The core genes are defined as genes found in all (100%) or a majority (e.g., 99%) of the studied genomes. Therefore, two parameters are critically important for the definition of core genes in pan-genome analysis: (i) in what percentage of genomes the core gene is found (e.g., 100% vs 99%); (ii) in searching each genome for the core gene what is the sequence identity (e.g., 95% vs 90%) used to call a presence. In all the above analyses, the core gene (CG) threshold was 100% and the sequence identity (SI) threshold was 90%. To determine the two parameters' impacts on pan-genome results, different CG thresholds and different SI thresholds were used to repeat all pan-genome analyses. For an *E. coli* dataset, the CG threshold varying from 100% to 85% were tested in different simulation groups by using Roary with a 90% SI threshold (**Figure 14 a, b, c**). In *E. coli* fragmentation groups shown in **Figure 14a**, a decreasing core genome size was observed with an increasing fragmentation when using core gene threshold 100%,

99%, and 98%. However, when using core gene threshold <=95%, the loss of core gene families caused by fragmentation was negligible.

In contrast, the effect of incompleteness is much larger than fragmentation. When using core gene threshold>=90%, a notable reduction in core genome size was observed in genomes with an average of 5% incompleteness (**Figure 14b**). When using core gene threshold even as low as 85% for genomes with average incompleteness >10% (the MIMAG recommended lowest incompleteness for high-quality MAGs), the core gene loss caused by incompleteness was still a problem. Similar results were also observed in *B. pertussis* fragmentation and incompleteness datasets (**Figure 15**). In contrast, contamination caused the smallest change in the core genome size irrespective of what core gene threshold is used (**Figure 14c**). All these suggest that incompleteness (or missing genes in MAGs) will significantly reduce the number of core genome size, and choosing a more relaxed CG threshold (e.g., > 90%) will help only a little and only when the incompleteness is less than 5%.

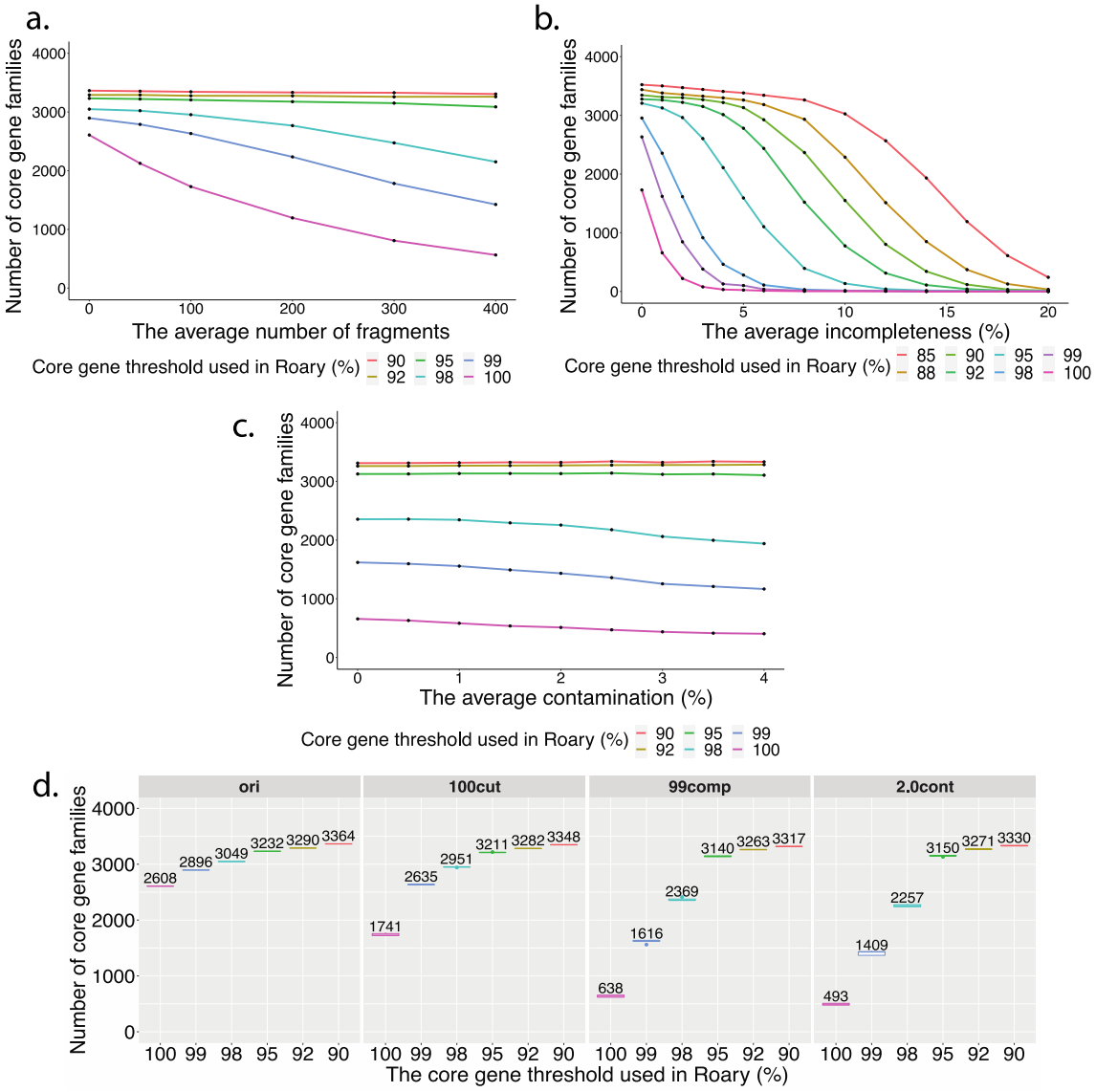

Figure 14. Different core gene thresholds in Roary influence the *E. coli* core genome. Line plots of the number of core gene families in *E. coli* datasets with different fragmentation (**a**), incompleteness (**b**), and intraspecies contamination(**c**) when using core gene threshold between 100% and 85%. Incompleteness groups simulated based on 100 fragmentation, and intraspecies contamination groups simulated from datasets having 100 fragmentation and 99% completeness. **d:** Boxplots of the number of core gene families in an *E. coli* original dataset and 5 *E. coli* simulation datasets with 100 fragmentation or 99%completeness or 2.0% contamination. Different colors represent different core gene thresholds. The numbers in "ori" group are core gene sizes in the original dataset. The number in "100cut", "99comp" and "2.0cont" groups are average core gene sizes for 5 simulation datasets.

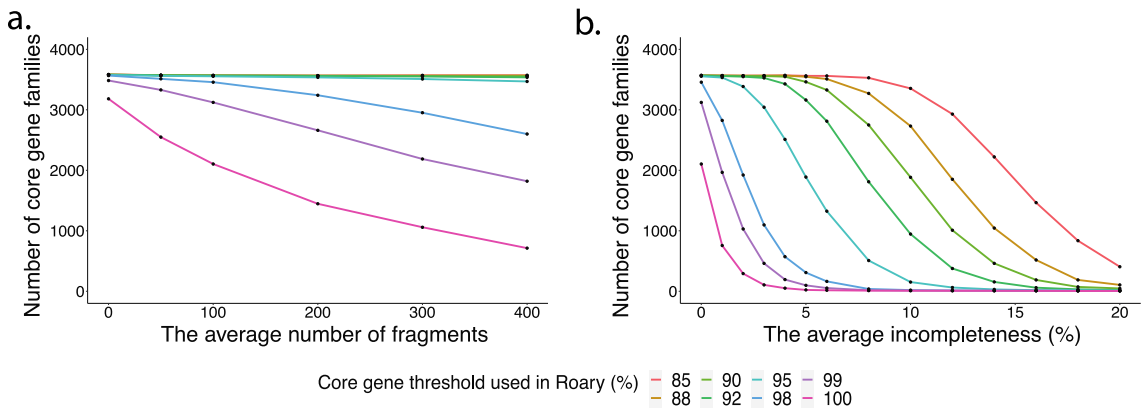

Figure 15. Different core gene thresholds in Roary influence the B. pertussis core genome. The line plots of the number of core gene families in *B. pertussis* datasets with different fragmentation (**a**), and incompleteness (**b**) when using core gene threshold between 100% and 85%. See legend in Figure 14.

We have also tested different core gene thresholds using 5 *E. coli* simulation datasets (an original dataset with 100 random-selected complete genomes and 5 times of simulation). The finding again revealed that the lower the core gene threshold was used, the closer the core gene size to that of the original genomes (**Figure 14d**). Overall, when using Roary for pan-genome analysis, the different core gene thresholds had significant effects on the number of core gene families in genomes having varying fragmentation and completeness.

The observations made in Roary were also true for BPGA (**Figure 16 a, b**). The core gene loss alleviation due to the core gene threshold was also observed in Anvi'o (**Figure 16d**). However, the core gene thresholds in Anvi'o had little effect on the number of core genes when increasing fragments in *E. coli* genomes (**Figure 16c**). It should be mentioned that Prodigal was used as the default gene annotation tool in Anvi'o, and the parameter differences between Prodigal in Anvi'o and Prokka (also calls Prodigal for gene prediction) used for BPGA and Roary may be different. Indeed, we noticed that

in Anvi'o the Prodigal gene prediction is run in metagenome mode, while in Prokka Prodigal gene prediction is run in normal mode (i.e., does not work for fragmented genes). This may explain why in Anvi'o, fragmentation had little effect on core gene loss.

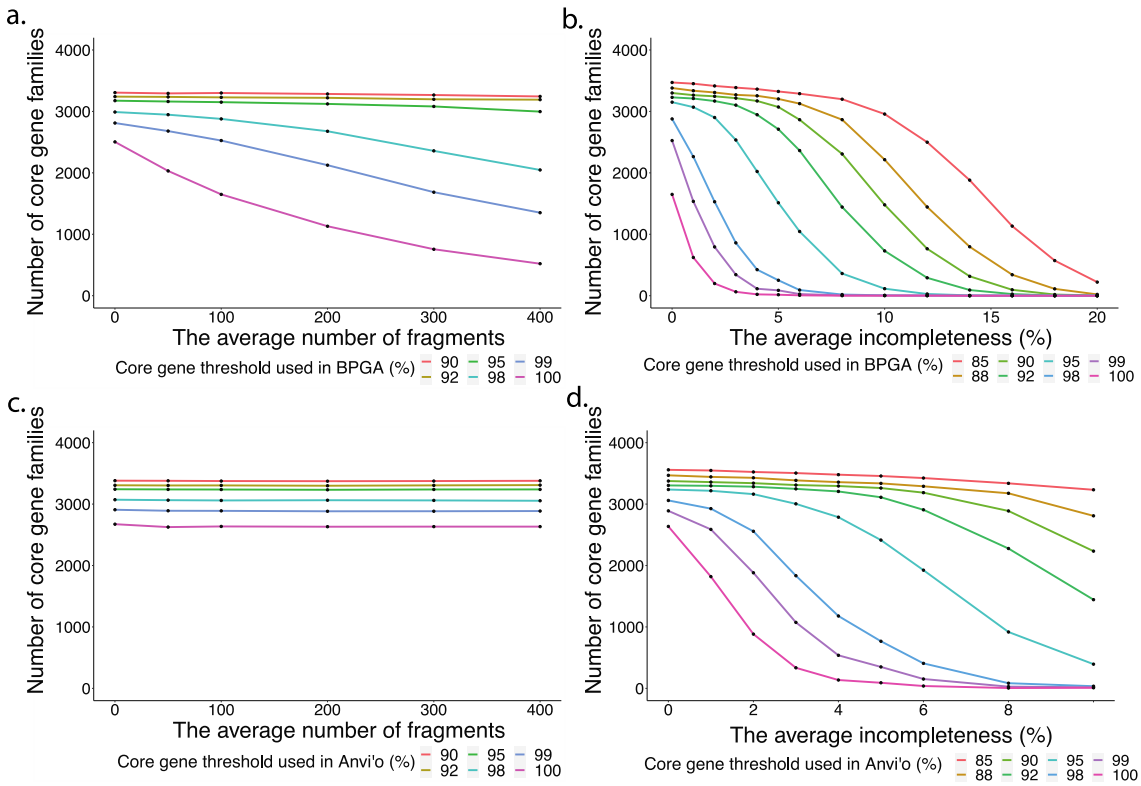

Figure 16. Different core gene thresholds in BPGA and Anvi'o influence the *E. coli* core genome. **a&b:** Line plots of the number of core gene families in *E. coli* datasets with different fragmentation (**a**), and incompleteness (**b**) when using core gene threshold between 100% and 85% in BPGA. **c&d:** Line plots of the number of core gene families in *E. coli* datasets with different fragmentation (**c**), and incompleteness (**d**) when using different core gene thresholds in Anvi'o.

Compared to the core gene threshold, the sequence identity threshold had very little effect on the core gene loss alleviation. Although various shapes of lines were observed from four figures shown in **Figure 17**, the lines representing different sequence

identities were overlapped irrespective of what core gene threshold was used (e.g., 100% core gene threshold in **Figure 17a**).

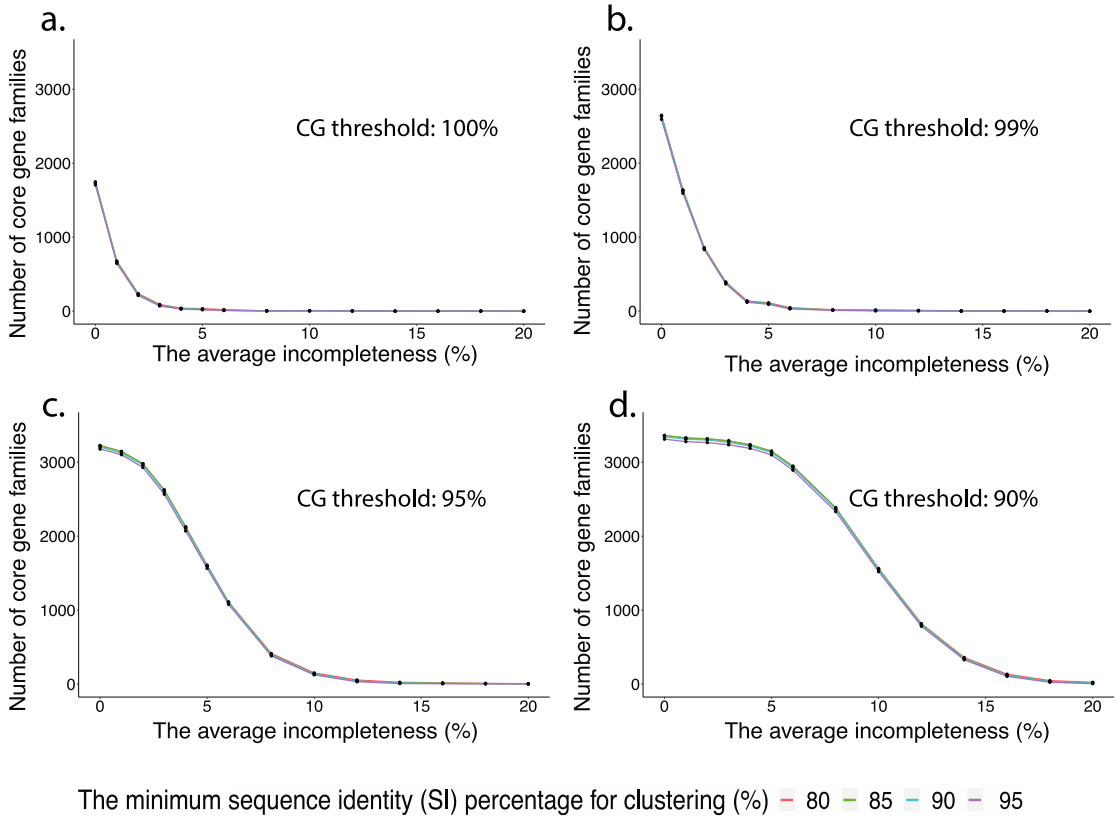

Figure 17. Different gene clustering identities in Roary. Line plots of the number of core gene families in *E. coli* datasets with different incompleteness percentages when using different clustering identities. The core gene (CG) threshold used is 100% (**a**), 99%(**b**), 95% (**c**) and 90% (**d**). Different colors represent the minimum identity percentage for gene clustering.

To conclude, the core gene threshold is a very important parameter. Choosing a more relaxed core gene threshold can help alleviate the core gene loss caused by genome fragmentation and incompleteness.

The decrease of core genes also leads to underestimation in core gene functional analysis

In pan-genome analysis, after the core and variable genes are identified, downstream analyses are often performed on these genes to better understand the evolution of different genomes and functions of different genes. Given that the core genome size is inevitably decreased in the pan-genome analysis of MAGs, there was some potential underestimation in the COG functional predictions for core genomes. When using the 100% core gene threshold, the number of core genes assigned to COG categories was dramatically decreased with increasing fragmentation or incompleteness. The number of core gene representatives in different *E. coli* pan-genome assigned to different COG categories were shown in **Figure 18**. Three COG categories ([E] Amino acid transport and metabolism, [G] Carbohydrate transport and metabolism, [J] Translation, ribosomal structure and biogenesis) were the most abundant in *E. coli* original core genomes. These COG categories were significantly changed with fragmentation, especially in genomes with more than 300 and 400 fragments (**Figure 18a**). Additionally, more than half of the core gene representatives were lost in each COG category when genomes have an average of 1% completeness loss (**Figure 18b**). Genes in some COG categories were completely lost when genomes had an average of 5% or 4% incompleteness. A similar observation was also made in *B. pertussis* (**Figure 19 a, b**).

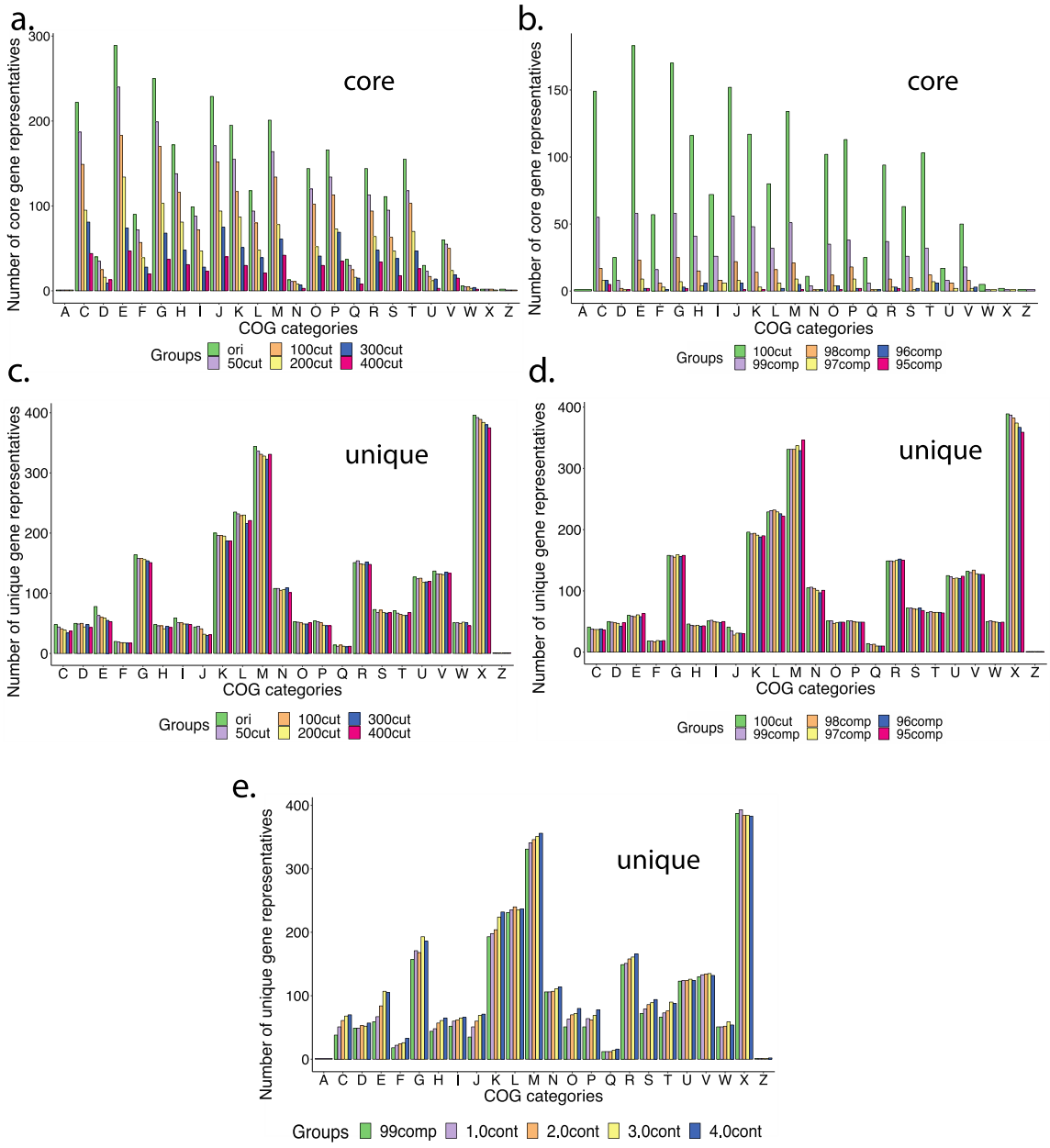

Figure 18. The COG analysis for core and unique gene representatives in E. coli. The bar plots of the number of core gene representatives in *E. coli* datasets with different fragmentation (**a**) and incompleteness (**b**) in each COG category. The bar plot of the number of unique gene representatives in *E. coli* datasets with different fragmentation (**c**), incompleteness (**d**), or intraspecies contamination (**e**) in each COG category. **COG categories**: A: RNA processing and modification, C: Energy production and conversion, D: Cell cycle control and mitosis, E: Amino Acid metabolism and transport, F: Nucleotide metabolism and transport, G: Carbohydrate metabolism and transport, H: Coenzyme metabolism, I: Lipid metabolism, J: Translation, K: Transcription, L: Replication and repair, M: Cell wall/membrane/envelop biogenesis, N: Cell motility, O: Post-translational modification, protein turnover, chaperone functions, P: Inorganic ion

transport and metabolism, Q: Secondary Structure, R: General Functional Prediction only, S: Function Unknown, T: Signal Transduction; U: Intracellular, trafficking and secretion, V: Defense mechanisms, W: Extracellular structures, X: Mobilome: prophages, transposons, Z: Cytoskeleton.

Compared to the core gene representatives, the number of *E. coli* unique gene (gene only found in 1 genome in the dataset) representatives in COG categories were slightly affected by fragmentation and incompleteness but changed by increasing intraspecies contamination (**Figure 18 c, d, e**). The addition of genomic contaminants from other strains may confound the overall functional prediction for unique genes. However, only 27%~33% of the unique gene representatives were associated with COG categories, the functions of a large proportion of unique genes are still unknown.

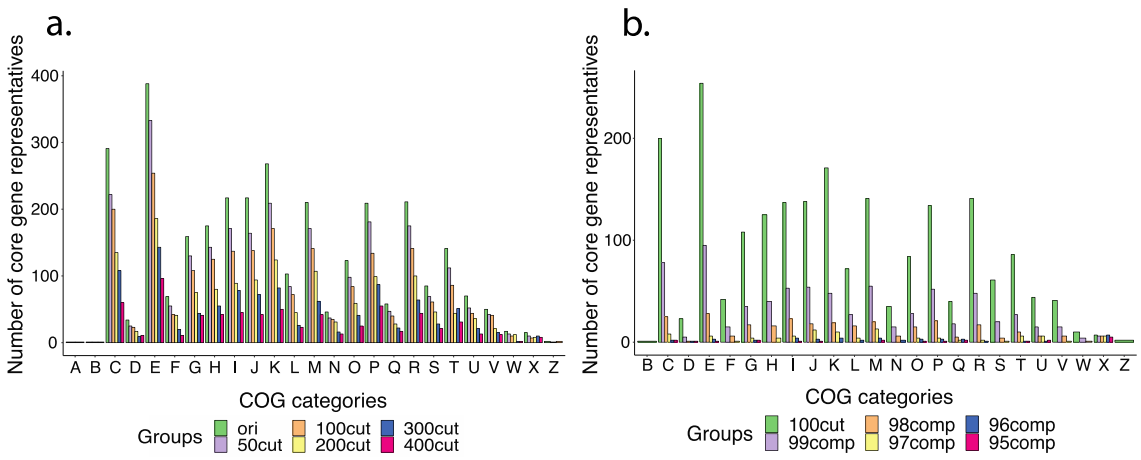

Figure 19. The COG analysis for core gene representatives in B. pertussis. The bar plots of the number of core gene representatives in *B. pertussis* datasets with different fragmentation (**a**) and incompleteness (**b**) in each COG category. See COG details in Figure 18.

Compared with using the 100% core gene threshold (**Figure 18 a, b**), more core genes will be kept when using a lower core gene threshold. Therefore, as expected, fewer decreases on core genome functional predictions caused by fragmentation were observed

when using the core gene threshold at 99%, 98% and 95% (**Figure 20 a, c, e**). When using 95% core gene threshold, few changes in the number of core gene representatives

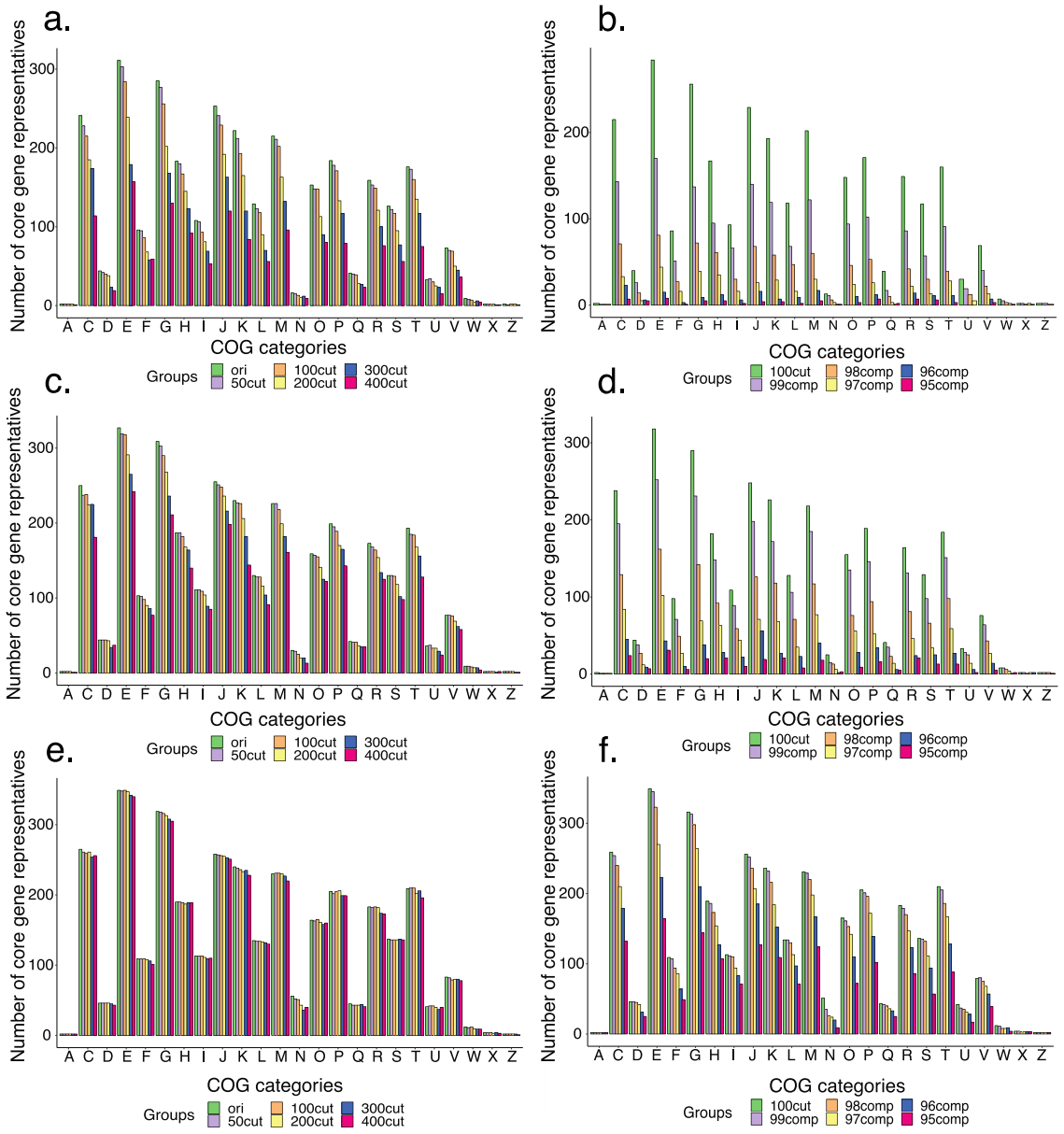

Figure 20. Core gene thresholds influence COG analysis in *E. coli* core genome. Bar plots of the number of core gene representatives in *E. coli* datasets with different fragmentation (left) and incompleteness (right) in each COG category. The core gene threshold 99% (**a b**), 98% (**c d**) and 95% (**e f**) are used.

in each COG category were due to the increasing fragments in genomes (**Figure 20e**). Similarly, fewer decreases on core genome functional predictions caused by incompleteness were observed with COG functions if a lower core gene threshold was used (**Figure 20 b, d, f**). Although the lower core gene threshold helped maintain more COG function predictions, the underestimation caused by fragmentation or incompleteness could not be eliminated. Even worse, the inclusion of non-core genes may lead to other misprediction in further analysis.

Phylogenetic trees are also affected by fragmentation, incompleteness, and contamination

From the pan-genome analysis result, a gene presence and absence matrix can be derived, which has rows representing all the genomes and columns representing the different gene clusters (or gene families), From this gene presence and absence matrix, a phylogenetic tree can be constructed to depict the evolutionary relationship among all the studied genomes in the pan-genome analysis. When the genomes were fragmented, incomplete and contaminated, the gene presence and absence matrix will likely change, and the phylogenetic tree will change as well. The tree of 100 original *E. coli* genomes (**Figure 21a)** was built by using the matrix of 17,961 gene clusters. A large dark blue area shown in the red frame represented core gene clusters. However, a dramatic shrinkage in the core gene area was observed in **Figure 21b** (simulated genomes with 100 fragments and 1% incompleteness). The white dots representing gene absence were evenly distributed in the accessory gene area in the green frame, indicating the gene loss in all the genomes. The overall tree topology changed a lot after 100 fragmentation and

99% completeness simulation, while some clades (e.g., the blue and purple ones in **Figure 21**) were more conserved than others.

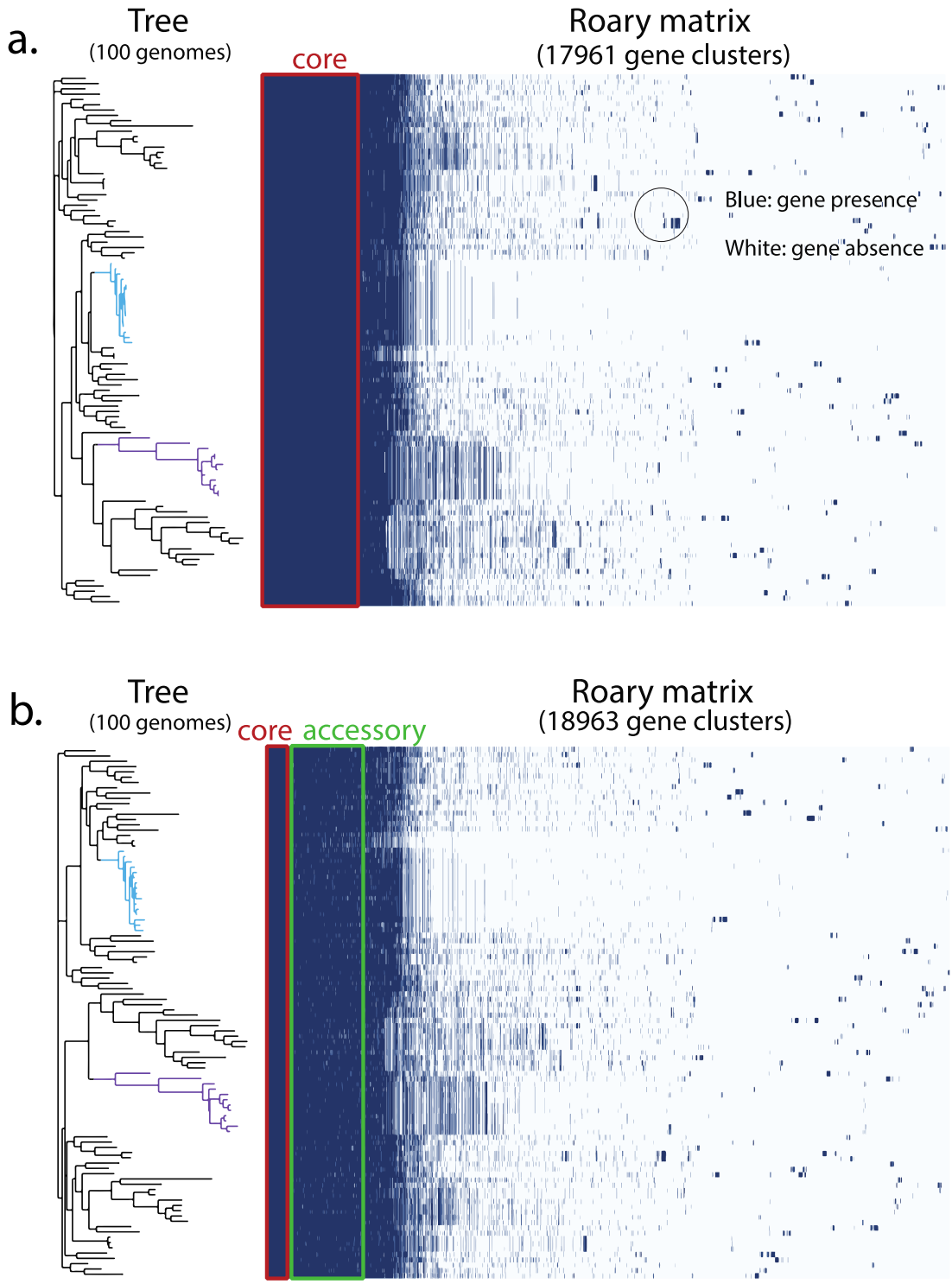

Figure 21. Phylogenetic trees of 100 *E. coli* genomes constructed by using Roary pan-genome results. **a**: The tree and Roary matrix for 100 original *E. coli* genomes. **b**: The tree and Roary matrix for 100 *E. coli* genomes with an average of 100 fragments and 99% completeness. Each row corresponds to a branch on the tree and represents one genome. Each column represents an orthologous gene family/cluster. White indicates gene absence and blue indicates gene presence in the gene cluster matrix.

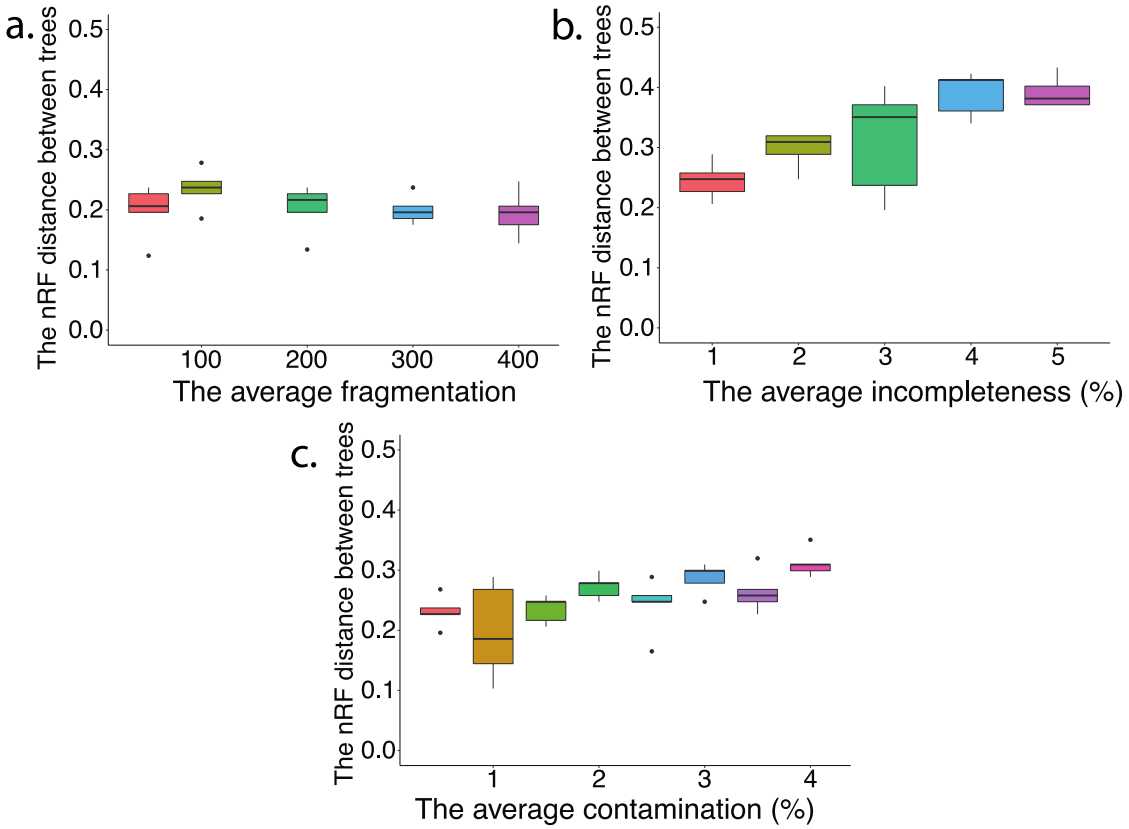

Figure 22. The nRF distance values between phylogenetic trees in five *E. coli* simulation datasets. The boxplots for nRF distance values between trees constructed for *E. coli* datasets with different fragmentation (**a**), incompleteness (**b**), and intraspecies contamination (**c**).

To more quantitively measure the difference between the trees before and after, Normalized Robinson-Foulds (nRF) distance was used. In five *E. coli* datasets, the nRF distance (Robinson and Foulds, 1981) between two phylogenetic trees increased continuously with more incompleteness in genomes, while nRF values slightly fluctuated

between 0.2 and 0.4 in fragmentation and contamination groups (**Figure 22**). Clearly, the loss of completeness had more effects on the phylogenetic analysis than fragmentation and contamination. Notable variations among datasets were found in some simulation groups (e.g. 3% incompleteness group in **Figure 22b** and 1% intraspecies contamination groups in **Figure 22c**). This might be caused by the very small number (5) of datasets used in this analysis.

A similar observation was made in the other two species, *S. aureus* and *B. pseudomallei* (**Figure 23**). The nRF distance values varied between 0 and 0.45 among 10 *S. aureus* datasets in different simulation groups (**Figure 23a**), however, the nRF distance values were larger than 0.9 in all the *B. pseudomallei* datasets (**Figure 23b**). The large nRF values indicated significant differences between phylogenetic trees constructed for complete genomes and simulated genomes, meaning likely impacts on the phylogeny study of *B. pseudomallei* when using MAGs with fragmentation, incompleteness, and contamination. The findings pointed out that the effects on phylogenetic analysis caused by the nature of MAGs may be different in different species.

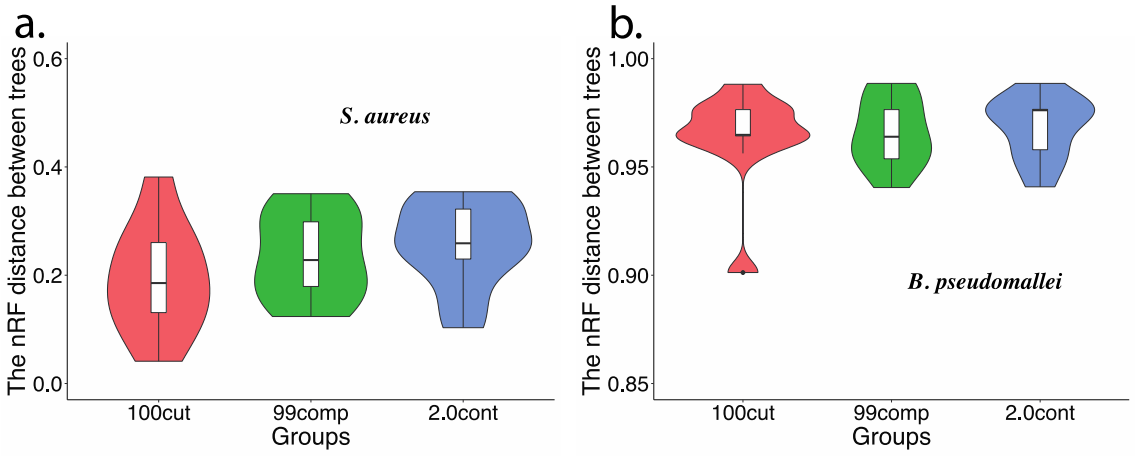

Figure 23. The nRF distance values in two species simulation datasets. Boxplots of the nRF distance values between trees in the simulation group and the original tree in *S. aureus* (**a**) and *B. pseudomallei* (**b**).

Simulation groups: 100cut represents an average of 100 fragments in genomes, 99comp represents an average of 99% completeness in genomes with 100cut, 2.0cont represents the genomes in datasets contain an average of 100 fragments, 99% completeness, and 2% intraspecies contamination. for each species, ten simulation datasets were used in each group.

In addition to the gene presence and absence matrix, phylogenetic trees can also be built based on core gene alignment. In five *E. coli* datasets simulated from an original 100-random genome dataset, the nRF distance values were lower than 0.2 when compared the original tree and the tree constructed using core gene alignment from genomes having 100 fragmentation (**Figure 24**). When using genomes with 1% incompleteness for tree construction, the nRF distance values between new and original trees increased to ~0.3. The nRF values for tree comparison were even higher than 0.4 when additional 2% intraspecies contamination was added to genomes. Therefore, there may be some bias or errors in phylogenetic analysis of species MAGs whether using gene presence/absence or core gene alignment for tree construction.

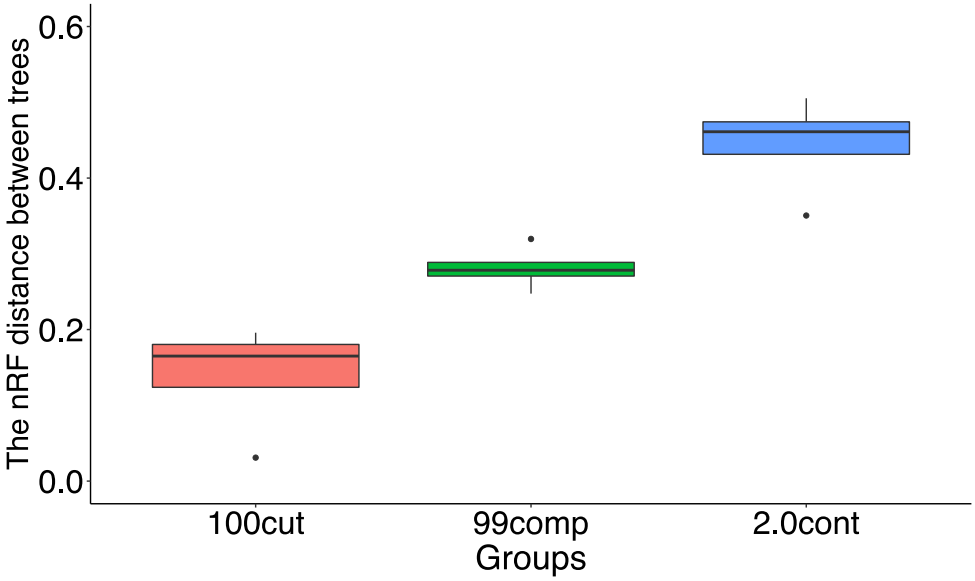

Figure 24. The nRF distance compares the core gene-based phylogenetic trees in 5 *E. coli* datasets.

See legend in **Figure 23**.

CHAPTER V

DISCUSSION AND CONCLUSION

1. DISCUSSION

17 Species Assessment and Their Pan-genomes

The 17 species used in the project are common bacterial pathogens. These species have a large number of complete genome sequences because they are well studied in clinical and environmental biology. More and more new isolates in these species are found and published in the NCBI RefSeq database (https://www.ncbi.nlm.nih.gov/pathogens/organisms). Since a single genome cannot reflect the entire genetic variability of a bacterial species (Costa *et al.*, 2020), a large number of genomes are desirable to be used in the pan-genome analysis (Tettelin *et al.*, 2005; Vernikos *et al.*, 2015; Costa *et al.*, 2020). A large number of genomes not only help to explore the species' genomic characteristics and variations (Park *et al.*, 2019) but also strengthen the expressive and statistical power of pan-genomic studies (Marschall *et al.*, 2018). Therefore, the 17 species, each of which contains at least 100 complete genomes, are chosen as the representative species for more accurate pan-genome analysis.

The differences in pan-genome size and structures among species reflected various genomic diversities and genetic characteristics. This finding can be explained and verified by considering ANI values, the number of plasmids in genomes, and the survival environment of bacteria. For example, the largest open pan-genome observed in *E. coli* indicated high diversity in different strain genomes. The large variation in pairwise ANI

values showed clear genomic differences within *E. coli* genomes. This is consistent with the previous finding that ANI values for *E. coli* genomes are distributed across a wide range, between 95.96 to 99.99% (Park *et al.*, 2019). Some *E. coli* genomes have even more than 10 plasmids, indicating that lots of adaptive genes related to antibiotic resistance, virulence, and metabolic adaptations may be contained to fit specific hosts and environmental conditions (de Toro *et al.*, 2014). Gene gain and loss in *E. coli* genomes lead to high variability in gene contents between lineages and isolates (Touchon *et al.*, 2009). The diversity of *E. coli* strains in various environments was revealed in some previous studies (Van Elsas *et al.*, 2011; Jang *et al.*, 2017). In contrast, *B. pertussis* has a large proportion of core genes within its small pan-genome. Since it is a human-specific bacterium and does not survive in the environment (Trainor *et al.*, 2015), the results including few accessory genes, no plasmid, and high ANI values reflect the low genomic plasticity and diversity in *B. pertussis*, which is consistent with previous findings (Mooi, 2010; Weigand *et al.*, 2017). Two species, *K. pneumoniae* and *S. enterica*, in the same family (*Enterobacteriaceae*) as *E. coli* have large pan-genome containing lots of accessory genes, indicating their substantial genetic diversity across diverse environmental niches (Holt *et al.*, 2015; Laing *et al.*, 2017). In contrast, *C. pseudotuberculosis* and *M. tuberculosis* belonging to *Actinobacteria* phylum have a smaller pan-genome size and a large proportion of core gene families, which is consistent with previous studies that they have close relationships among strains (Soares *et al.*, 2013; Dar *et al.*, 2020). Compared to *M. tuberculosis*, some *C. pseudotuberculosis* strains have more variable genes gained through horizontal gene transfer (Soares *et al.*, 2013), which is consistent with ANI value differences between genomes.

The pan-genomes of all the 17 species are determined as “open”. The results are consistent with previous findings (Qin *et al.*, 2012; Chan *et al.*, 2015; Spring-Pearson *et al.*, 2015; van Vliet, 2017; Lu *et al.*, 2019; Costa *et al.*, 2020). However, compared to other species, *B. pertussis* has its pan-genome near to be closed according to a large percentage of core genes and less genomic plasticity (Costa *et al.*, 2020). It should be mentioned that all the pan-genomes were built using only complete genomes. Including draft genomes from RefSeq may strengthen the finding that all the species have an open pan-genome.

Core Genome Loss is most affected by Incompleteness, followed by Fragmentation and contamination

For all 17 species used in this study, the core genome loss was observed when using simulated MAGs with different levels of fragmentation and incompleteness. These results are expected given that genomes with missing gene fragments would lead to reduced core genome size in pan-genome analysis (Zhou *et al.*, 2020). We have illustrated the genomic consequences of having simulated fragmentation, incompleteness, and contamination in a given genome (**Figure 25)**.

First, for fragmentation, if the randomly selected cut position is within a gene coding region, the open reading frame (ORF) will be split into two fragments. Some ORFs may still be predicted as genes or as hypothetical proteins, while others may be completely lost due to the missing of start or stop codons. Prokka (Seemann, 2014) automatically uses Prodigal (Hyatt *et al.*, 2010) with a “closed ends” flag (do not allow genes to run off edges.) for gene annotation and prediction. Therefore, only full-length

genes can be predicted, and partial/fragmented genes are not predicted. The sequences cut at the beginning of the ORFs may lead to the loss of the first start codon, the ORFs may start at the second/third start codon and lead to completely different protein predictions (Sarkar *et al.*, 2019; Huang *et al.*, 2021).

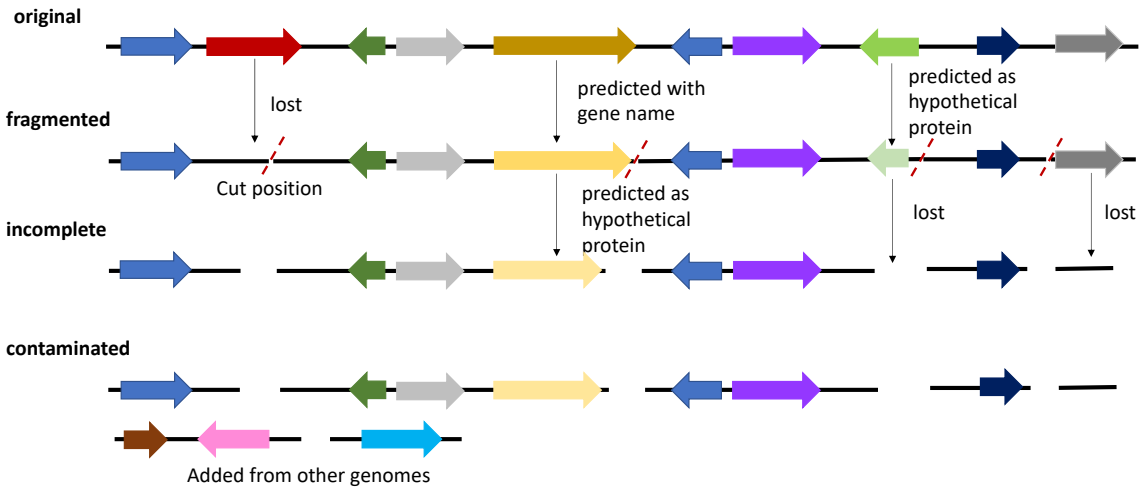

Figure 25. The diagram showing gene loss in genome fragmentation and incompleteness simulation. Black straight lines represent genome sequences. Arrows shown in different colors represent predicted genes (or open reading frames) in genome sequences.

Second, to simulate incompleteness, some ORFs are partially or fully removed when creating gaps between genomic fragments. Therefore, the gene prediction step is more significantly affected, leading to missing genes. Even if there are some genes predicted in the two ends of genomic fragments, they might be coded into totally different proteins due to reading frame shifts, leading to falsely predicted genes (Lomsadze *et al.*, 2018). Moreover, due to a large number of genomes analyzed, only a few errors in gene prediction can significantly change pan-genome analysis results (Costa *et al.*, 2020).

Third, to simulate contamination, sequence fragments from closely related strains or species are added. Unlike fragmentation and incompleteness, there is no removal of existing genomic regions. Therefore, there should be no loss of core genes. However, the size of the core genome still decreased when adding intra- or interspecies contamination (**Figure 10 and Figure 11**) when using Roary. In contrast, the core genome size did not decrease in BPGA and Anvi'o (Figure 13). This raised the question that if Roary has limitations to deal with contaminations.

We have looked further into Roary's (Page *et al.*, 2015) analysis result. It appears that some core gene clusters before the contamination are falsely split into multiple gene clusters (**Appendix B**). As a result, some core genes were no longer considered as core genes due to the clustering error, leading to the underestimation of core genome size. Gene clusters being incorrectly split into multiple smaller clusters were also noticed in other studies (Tonkin-Hill *et al.*, 2020; Zhou *et al.*, 2020). Interestingly, the addition of contamination sequences from closely related strains of the same species will lead to more errors in gene clustering than contamination from other species.

Tonkin-Hill *et al.* (2020) also evaluated the effects of errors on pan-genome analysis caused by annotation errors, fragmented assemblies, and contamination in their recent study; the results of core genome reduction are consistent with our observation. They suggested that the removal of contamination and correction of annotation errors are essential to construct an accurate pan-genome for genomes with fragmentation and contamination (Tonkin-Hill *et al.*, 2020). Compared to their simulated genomes focusing on genome fragmentation, contamination, and gene gain/loss in the accessory genome, our study also simulated genomes to represent incompleteness in real MAGs. Our results

indicated that genome completeness was also important for accurate pan-genome analysis.

Although the core genome reduction was observed in all the 17 species (**Figure 8 and Figure 9**), the different rates of core genome loss may be related to the characteristic of species. In general, species that have higher ANI values will have larger core genomes because of the close relationships among strains. These species tend to have a more rapid reduction in core genome sizes. Since the core genome reduction varied among species, it is still difficult to predict the loss of core gene families in real MAGs.

Different Performances in Roary, BPGA, and Anvi'o

For both *E. coli* and *B. pertussis* datasets, the core genome size predicted by using Roary and BPGA is consistent, while different predictions are given in the contamination groups (**Figure 13**). The different performance is possibly correlated with the algorithm used for gene clustering. In Roary, CD-HIT (Li and Godzik, 2006) is firstly used to reduce sequence redundancy and select the representative sequences. The longest input sequence is picked as the first cluster representative, the remaining sequences from long to short are compared based on their similarities to the existing representatives. The sequences will be classified as a redundant or representative sequence (Fu *et al.*, 2012). Then MCL (Van Dongen and Abreu-Goodger, 2012) is used to cluster the representative sequences. Since only a single set of representatives is selected for clustering, it is a critical limitation that this set may not be the best representation for the whole dataset (Surujonu *et al.*, 2020). If the representative (longest) sequence is selected from the contamination sequences, the clustering results may be changed consequently. In

comparison, Usearch (Edgar, 2010) is used as the clustering tool for BPGA. Target sequences are compared to the query in order of decreasing unique word count (U), and sequences above the identity threshold will be considered as a hit. If a hit exists, it will be found among the first few candidates in the sorted U-list. Therefore, it has a higher speed and more improved sensitivity than CD-HIT (Edgar, 2010).

Roary and BPGA have more similar results in *B. pertussis* than in *E. coli*, indicating that their clustering performance will be influenced by different species. Due to possible effects on clustering caused by a large number of accessory genes in *E. coli*, these pan-genome tools may have a more accurate result for species like *B. pertussis*, which have more conserved core genomes, than *E. coli*. This finding is consistent with that performance of pan-genome computational tools decreased with increasing levels of genome variations and evolutionary distance (Bonnici *et al.*, 2021).

The performance of Roary and Anvi'o changes when using different gene prediction tools. When the gene annotation provided by Prokka is used, the core genome size predicted by Roary and Anvi'o is similar. However, if the default Prodigal in the Anvi'o pipeline is used, less core genome reduction will be caused by fragmentation (**Figure 14a and Figure 16a**). In Anvi'o, the prodigal is used with "-p meta" for metagenome mode as default if no translation tables are given; it also predicts genes that run off the edges of the sequence. In other words, more fragmented/partial genes will be predicted in Prodigal than that in Prokka. However, in Prokka, the candidate genes will be searched against domain-specific databases and identified with models of protein families, the gene prediction results may be fewer but more accurate. Therefore, it is still

unclear whether using Anvi'o with its default Prodigal is a better choice for more accurate pan-genome analysis on MAGs.

Important Parameters for Pan-genome Analysis

In pan-genome analysis, the lower the core gene threshold is used, the larger the core genome size is predicted (**Figure 14, Figure 15, and Figure 16**). However, when using a lower core gene threshold (e.g., 80%), some gene clusters that should be classified as accessory gene families may be falsely considered as part of the core genome, leading to possible overestimation in essential genes for microbial survival. In contrast, the strict core gene threshold (genes present in 100% or >99% of genomes) is not suggested due to the fragmentation, incompleteness and contamination in MAGs. The core gene threshold used in the pan-genome analysis for MAGs should be carefully selected based on the quality of MAGs. Therefore, there should be a good balance between the size and accuracy of the core genomes. For instance, a higher core gene threshold like 95% should be applied to MAGs with only low fragmentation (average number of fragments<100) and/or low incompleteness (average incompleteness<5%), while a lower core gene threshold like 90% or 85% should be used for MAGs with lots of fragmentation (average number of fragments>100), and/or high incompleteness (average incompleteness>5%).

When the identity for gene clustering is higher than 80%, the size of the core genome predicted by using Roary was dramatically influenced (**Figure 17**). Since the clustering identity higher than 90% is suggested in Roary, the identity lower than 80% is not evaluated in this study. In Anvi'o pan-genome analysis, the MCL inflation, which

defines the sensitivity of the MCL algorithm in the identification of the gene clusters, is used instead of sequence identity. The MCL inflation is set as "10" for comparing strains in species (Eren *et al.*, 2015). Chan *et al.* evaluated the clustering accuracy of MCL and UCLUST (share the same algorithm with Usearch) using different inflation values and clustering identity. They found that factors including sequence divergence and GC content bias would affect the accuracy of sequence clustering (Chan *et al.*, 2013). The appropriate clustering identity or MCL inflation should be selected for MAGs based on their quality (e.g., incompleteness or contamination percentages). For the high-quality MAGs, clustering identity higher than 80% or MCL inflation of 10 is suggested. If MAGs are more incomplete and contaminated, clustering identity 50%-80% or MCL inflation 2-10 should be tested to find the most suitable parameters.

In studies that use pan-genome analysis on MAGs (Appendix: Supplemental 1), it is noticed that the core gene threshold varies from 66% to 100%. The clustering identity used in Roary and BPGA ranged between 50% and 95%, while the MCL inflation values used in Anvi'o were 2, 5, and 10. Further detailed guidance and evaluations are needed to help select the parameters used in the pan-genome analysis for MAGs.

Underestimation and Misprediction in Downstream Analysis of MAGs

The reduction in core genome size could naturally lead to the underestimation of the COG functions for core genomes and the misprediction in phylogenetic trees. Although using a lower core gene threshold (e.g., 95%) may maintain more COG functions (**Figure 20**), some functional prediction errors may be caused by including the misclassified core gene representatives. In previous studies of MAGs, the COG

functional analysis was performed to study the enriched COG functional categories in core and accessory genes of all species found in the human microbiome (Almeida *et al.*, 2021), to determine functional differences between gene clusters in the *Sulfurovum* pangenomes (Moulana *et al.*, 2020), and to investigate key functions shared by genomes in a phylum (Shaiber *et al.*, 2020). Therefore, the COG analysis for core genes may help to understand the survival and adaptive ability of species, while the function studies for unique genes are especially important to understand the differences among strains in various environments. For MAG studies, the accuracy of core/unique genomes is the key for the precise interpretation of functions in MAGs, so the pan-genome analysis needs to be improved in the future to ensure more accurate core/unique genomes for MAGs.

The errors in phylogenetic trees will lead to the misinterpretation of the species evolution and relationships among strains. For using phylogenetic trees built from gene presence and absence, it has been reported that the lost core genes in some strains will affect the inference of gene gain and acquisition based on the phylogeny (Gabrielaite and Marvig, 2020). The accuracy of these phylogenetic trees highly depends on the completeness and fragmentation of genomes (Gabrielaite and Marvig, 2020). Our data clearly showed that phylogenetic trees based on gene presence and absence will be much more likely to be affected when it is based on MAGs than when based on complete genomes. The accuracy of phylogenetic trees based on core genome alignment will be also significantly influenced by the reduction of core genomes. The genomic features including single nucleotide polymorphisms (SNPs) may be lost with core genome shrinkage, leading to inaccurate strain relationships.

Limitations and Bias

There are some limitations and biases in this study. Firstly, all the 17 bacterial species are pathogenic. The lack of non-pathogenic species may not reveal all the issues in pan-genome accuracy and leads to possible bias in understanding the pan-genome accuracy loss. Secondly, the simulated MAGs may not fully represent the assembly and binning errors in real MAGs. In addition, some pan-genome analyses of MAGs may also include complete and draft isolate genomes. Our simulated genomes only used MAG-like genomes may lead to an overconcern of the problem. Moreover, the normalized Robinson-Foulds symmetric distances (nRFs) used to determine the differences between two phylogenetic trees can not consider the similarities between two clades, leading to possible errors in prediction. Additionally, only three pan-genome computational tools used in this study may not reveal the common problems in all tools. For example, the Panaroo pipeline(Tonkin-Hill *et al.*, 2020), which is designed to prevent the problems of low-quality draft genomes, is not evaluated in this study.

2. CONCLUSION

In this study, the differences of pan-genome analysis on complete genomes and simulated MAGs were compared among 17 species. A Python simulation pipeline was developed to generate MAGs with different levels of fragmentation, incompleteness and contamination from complete genomes. The incompleteness and fragmentation were the two most important reasons for core genome loss in MAGs, while the contamination influenced the number of unique gene families. The core genome reductions would

further affect the downstream analysis, leading to underestimation in COG functional prediction and misprediction in phylogenetic trees.

To improve the accuracy of pan-genome analysis on MAGs, the quality control of MAGs is indispensable as the first step. Although a new pipeline like panaroo has been developed to solve the errors caused by fragmentation and contamination in prokaryotic genomes, the tool is not designed for MAGs. The gene annotation and gene clustering algorithms considering the characteristics of MAGs (e.g., incompleteness) need to be developed to improve the core genome prediction. Moreover, the two important parameters, core gene threshold and clustering identity should be selected appropriately based on the quality of MAGs to make a balance between underestimation and overestimation of core genomes.

Overall, this study filled the research gaps in evaluating pan-genome accuracy on MAGs, revealed possible issues in studying pan-genome of MAGs, and provided suggestions on improving the pan-genome accuracy. The more accurate pan-genome analysis on MAGs will significantly improve the studies in the human gut microbiome, the spread and evolution of foodborne diseases, and the virulence-associated genes in pathogens.

REFERENCES:

Almeida,A. *et al.* (2019) A new genomic blueprint of the human gut microbiota. *Nature*, **568**, 499–504.

Almeida,A. *et al.* (2021) A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.*, **39**, 105–114.

Alneberg,J. *et al.* (2014) Binning metagenomic contigs by coverage and composition. *Nat. Methods*, **11**, 1144-1146.

Alneberg,J. *et al.* (2018) Genomes from uncultivated prokaryotes: A comparison of metagenome-assembled and single-amplified genomes. *Microbiome*, **6**, 173.

Anani,H. *et al.* (2020) Interest of bacterial pangenome analyses in clinical microbiology. *Microb. Pathog.*, **149**, 104275.

Anderson,R.E. *et al.* (2017) Genomic variation in microbial populations inhabiting the marine subseafloor at deep-sea hydrothermal vents. *Nat. Commun.*, **8**, 1114.

Baker,J.L. *et al.* (2021) Deep metagenomics examines the oral microbiome during dental caries, revealing novel taxa and co-occurrences with host molecules. *Genome Res.*, **31**, 64-74.

Becraft,E.D. *et al.* (2017) Rokubacteria: Genomic giants among the uncultured bacterial phyla. *Front. Microbiol.*, **8**, 2264.

Benedict,M.N. *et al.* (2014) ITEP: An integrated toolkit for exploration of microbial pan-genomes. *BMC Genomics*, **15**, 8.

Bezuidt,O.K. *et al.* (2016) The Geobacillus pan-genome: Implications for the evolution of the genus. *Front. Microbiol.*, **7**, 723.

Bharti,R. and Grimm,D.G. (2021) Current challenges and best-practice protocols for

microbiome analysis. *Brief. Bioinform.*, **22**, 178-193.

Bonnici,V. *et al.* (2021) Challenges in gene-oriented approaches for pangenome content discovery. *Brief. Bioinform.*, **22**, bbaa198.

Bowers,R.M. *et al.* (2017) Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.*, **35**, 725-731.

Brown,C.T. *et al.* (2015) Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature*, **523**, 208-211.

Buchanan,C.J. *et al.* (2017) A genome-wide association study to identify diagnostic markers for human pathogenic campylobacter jejuni strains. *Front. Microbiol.*, **8**, 1224.

Buchfink,B. *et al.* (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.

Carlos Guimaraes,L. *et al.* (2015) Inside the Pan-genome - Methods and Software Overview. *Curr. Genomics*, **16**, 245-252.

Castresana,J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.*, **17**, 540-552.

Chan,A.P. *et al.* (2015) A novel method of consensus pan-chromosome assembly and large-scale comparative analysis reveal the highly flexible pan-genome of Acinetobacter baumannii. *Genome Biol.*, **16**, 143.

Chan,C.X. *et al.* (2013) Clustering evolving proteins into homologous families. *BMC Bioinformatics*, **14**, 120.

Chaudhari,N.M. *et al.* (2016) BPGA-an ultra-fast pan-genome analysis pipeline. *Sci.*

Rep., **6**, 24373.

Chen,C. *et al.* (2021) Expanded catalog of microbial genes and metagenome-assembled genomes from the pig gut microbiome. *Nat. Commun.*, **12**, 1160.

Chen,L. *et al.* (2020) Accurate and complete genomes from metagenomes. *Genome Res.*, **30**, 315-333.

Chu,C. *et al.* (2019) GAPPadder: A sensitive approach for closing gaps on draft genomes with short sequence reads. *BMC Genomics*, **20**, 315-333.

Conlan,S. *et al.* (2012) Staphylococcus epidermidis pan-genome sequence analysis reveals diversity of skin commensal and hospital infection-associated isolates. *Genome Biol.*, **13**, R64.

Contreras-Moreira,B. and Vinuesa,P. (2013) GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Appl. Environ. Microbiol.*, **79**, 7696-7701.

Costa,S.S. *et al.* (2020) First Steps in the Analysis of Prokaryotic Pan-Genomes. *Bioinform. Biol. Insights*, **14**, 1177932220938064.

Dar,H.A. *et al.* (2020) Pangenome analysis of mycobacterium tuberculosis reveals core-drug targets and screening of promising lead compounds for drug discovery. *Antibiotics*, **9**, 819.

Delmont,T.O. *et al.* (2011) Accessing the soil metagenome for studies of microbial diversity. *Appl. Environ. Microbiol.*, **77**, 1315-1324.

Ding,W. *et al.* (2018) panX: pan-genome analysis and exploration. *Nucleic Acids Res.*, **46**, e5.

Van Dongen,S. and Abreu-Goodger,C. (2012) Using MCL to extract clusters from

networks. *Methods Mol. Biol.*, **804**, 281-295.

Edgar,R.C. (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792-1797.

Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460-2461.

Van Elsas,J.D. *et al.* (2011) Survival of Escherichia coli in the environment: Fundamental and public health aspects. *ISME J.*, **5**, 173-183.

Emms,D.M. and Kelly,S. (2019) OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.*, **20**, 238.

Eren,A.M. *et al.* (2015) Anvi'o: An advanced analysis and visualization platformfor 'omics data. *PeerJ*, **3**, e1319.

De Filippis,F. *et al.* (2020) Newly Explored Faecalibacterium Diversity Is Connected to Age, Lifestyle, Geography, and Disease. *Curr. Biol.*, **30**, 4932-4943.

Freschi,L. *et al.* (2019) The Pseudomonas aeruginosa Pan-Genome Provides New Insights on Its Population Structure, Horizontal Gene Transfer, and Pathogenicity. *Genome Biol. Evol.*, **11**, 109-120.

Fu,L. *et al.* (2012) CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150-3152.

Gabrielaite,M. and Marvig,R.L. (2020) GenAPI: A tool for gene absence-presence identification in fragmented bacterial genome sequences. *BMC Bioinformatics*, **21**, 320.

Galperin,M.Y. *et al.* (2021) COG database update: Focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.*, **49**, D274-D281.

Garcia,S.L. *et al.* (2015) Auxotrophy and intrapopulation complementary in the interactome of a cultivated freshwater model community. *Mol. Ecol.*, **24**, 4449-4459.

Georgiades,K. and Raoult,D. (2011) Defining pathogenic bacterial species in the genomic era. *Front. Microbiol.*, **1**, 151.

Handelsman,J. *et al.* (1998) Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chem. Biol.*, **5**, R245-249.

Higgins,P.G. *et al.* (2017) Development and evaluation of a core genome multilocus typing scheme for whole-genome sequence-based typing of Acinetobacter baumannii. *PLoS One*, **12**, e0179228.

Holt,K.E. *et al.* (2015) Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae, an urgent threat to public health. *Proc. Natl. Acad. Sci. U. S. A.*, **112**, E3574-3581.

Huang,X. *et al.* (2021) Frame-shifted proteins of a given gene retain the same function. *Nucleic Acids Res.*, **48**, 4396-4404.

Hugerth,L.W. *et al.* (2015) Metagenome-assembled genomes uncover a global brackish microbiome. *Genome Biol.*, **16**, 279.

Hyatt,D. *et al.* (2010) Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.

Jain,C. *et al.* (2018) High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.*, **9**, 5114.

Jang,J. *et al.* (2017) Environmental Escherichia coli: ecology and public health implications—a review. *J. Appl. Microbiol.*, **123**, 570-581.

Jégousse,C. *et al.* (2021) A total of 219 metagenome-assembled genomes of

microorganisms from Icelandic marine waters. *PeerJ*, **9**, e11112.

Kaas,R.S. *et al.* (2012) Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse Escherichia coli genomes. *BMC Genomics*, **13**, 577.

Kang,D.D. *et al.* (2015) MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, **2015**, e1165.

Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.

Kim,Y. *et al.* (2020) Current status of pan-genome analysis for pathogenic bacteria. *Curr. Opin. Biotechnol.*, **63**, 54-62.

Konstantinidis,K.T. and Tiedje,J.M. (2005) Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 2567-2572.

Kroeger,M.E. *et al.* (2018) New biological insights into how deforestation in amazonia affects soil microbial communities using metagenomics and metagenome-assembled genomes. *Front. Microbiol.*, **9**, 1635.

Laing,C.R. *et al.* (2017) Pan-genome analyses of the species Salmonella enterica, and identification of genomic markers predictive for species, subspecies, and serovar. *Front. Microbiol.*, **8**, 1345.

Letunic,I. and Bork,P. (2021) Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* gkab301.

Li,L. *et al.* (2003) OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178-2189.

Li,W. and Godzik,A. (2006) Cd-hit: A fast program for clustering and comparing large

sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658-1659.

Livingstone,P.G. *et al.* (2018) Genome Sequencing and Pan-Genome Analysis of 23 Corallococcus spp. Strains Reveal Unexpected Diversity, With Particular Plasticity of Predatory Gene Sets. *Front. Microbiol.*, **9**, 3187.

Lomsadze,A. *et al.* (2018) Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes. *Genome Res.*, **28**, 1079-1089.

Lu,Q.F. *et al.* (2019) Genus-wide comparative genomics analysis of Neisseria to identify new genes associated with pathogenicity and niche adaptation of Neisseria pathogens. *Int. J. Genomics*. 6015730.

Lu,S. *et al.* (2020) CDD/SPARCLE: The conserved domain database in 2020. *Nucleic Acids Res.*, **48**, D265-D268.

Marschall,T. *et al.* (2018) Computational pan-genomics: Status, promises and challenges. *Brief. Bioinform.*, **19**, 118-135.

Martin-Cuadrado,A.B. *et al.* (2008) Hindsight in the relative abundance, metabolic potential and genome dynamics of uncultivated marine archaea from comparative metagenomic analyses of bathypelagic plankton of different oceanic regions. *ISME J.*, **2**, 865-886.

Medini,D. *et al.* (2005) The microbial pan-genome. *Curr. Opin. Genet. Dev.*, **15**, 589–594.

Meyer,J.L. *et al.* (2017) Comparative metagenomics of the polymicrobial black band disease of corals. *Front. Microbiol.*, **8**, 618.

Mineeva,O. *et al.* (2020) DeepMAsED: Evaluating the quality of metagenomic assemblies. *Bioinformatics*, **36**, 3011-3017.

Mooi,F.R. (2010) Bordetella pertussis and vaccination: The persistence of a genetically monomorphic pathogen. *Infect. Genet. Evol.*, **10**, 36-49.

Moulana,A. *et al.* (2020) Selection Is a Significant Driver of Gene Gain and Loss in the Pangenome of the Bacterial Genus Sulfurovum in Geographically Distinct Deep-Sea Hydrothermal Vents. *mSystems*, **5**, e00673-19.

Nayfach,S. *et al.* (2019) New insights from uncultivated genomes of the global human gut microbiome. *Nature*, **568**, 505-510.

Nelson,W.C. and Stegen,J.C. (2015) The reduced genomes of Parcubacteria (OD1) contain signatures of a symbiotic lifestyle. *Front. Microbiol.*, **6**, 713.

O'Callaghan,A. *et al.* (2015) Pangenome analysis of Bifidobacterium longum and site-directed mutagenesis through by-pass of restriction-modification systems. *BMC Genomics*, **16**, 832.

O'Leary,N.A. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.

Olson,N.D. *et al.* (2019) Metagenomic assembly through the lens of validation: Recent advances in assessing and improving the quality of genomes assembled from metagenomes. *Brief. Bioinform.*, **20**, 1140–1150.

Page,A.J. *et al.* (2015) Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, **31**, 3691-3693.

Pantoja,Y. *et al.* (2020) Bioinformatics approaches applied in pan-genomics and their challenges. In, *Pan-genomics: Applications, Challenges, and Future Prospects*, pp. 43-64.

Park,S.C. *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**, 834.

Parks,D.H. *et al.* (2015) CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.*, **25**, 1043-1055.

Pasolli,E. *et al.* (2019) Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell*, **176**, 649-662.

Peng,X. *et al.* (2021) Genomic and functional analyses of fungal and bacterial consortia that enable lignocellulose breakdown in goat gut microbiomes. *Nat. Microbiol.*, **6**, 499-511.

Price,M.N. *et al.* (2010) FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.

Qin,X. *et al.* (2012) Complete genome sequence of Enterococcus faecium strain TX16 and comparative genomic analysis of Enterococcus faecium genomes. *BMC Microbiol.*, **12**, 135.

Quince,C. *et al.* (2017) Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.*, **35**, 833-844.

Reveillaud,J. *et al.* (2019) The Wolbachia mobilome in Culex pipiens includes a putative plasmid. *Nat. Commun.*, **10**, 1051.

Richter,M. and Rosselló-Móra,R. (2009) Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 19126-19131.

Riesenfeld,C.S. *et al.* (2004) Metagenomics: Genomic analysis of microbial

communities. *Annu. Rev. Genet.*, **38**, 525-552.

Robinson,D.F. and Foulds,L.R. (1981) Comparison of phylogenetic trees. *Math. Biosci.*, **53**, 131-147.

Roisin,S. *et al.* (2016) Pan-genome multilocus sequence typing and outbreak-specific reference-based single nucleotide polymorphism analysis to resolve two concurrent Staphylococcus aureus outbreaks in neonatal services. *Clin. Microbiol. Infect.*, **22**, 520-526.

Saghaï,A. *et al.* (2015) Metagenome-based diversity analyses suggest a significant contribution of non-cyanobacterial lineages to carbonate precipitation in modern microbialites. *Front. Microbiol.*, **6**, 797.

Sangwan,N. *et al.* (2016) Recovering complete and draft population genomes from metagenome datasets. *Microbiome*, **4**, 8.

Sarkar,K. *et al.* (2019) A frame-shifted gene, which rescued its function by non-natural start codons and its application in constructing synthetic gene circuits. *J. Biol. Eng.*, **13**, 20.

Schloss,P.D. and Handelsman,J. (2008) A statistical toolbox for metagenomics: Assessing functional diversity in microbial communities. *BMC Bioinformatics*, **9**, 34.

Seemann,T. (2014) Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068-2069.

Shaiber,A. *et al.* (2020) Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. *Genome Biol.*, **21**, 292.

Sheridan,P.O. *et al.* (2020) Gene duplication drives genome expansion in a major lineage

of Thaumarchaeota. *Nat. Commun.*, **11**, 5494.

Simon,C. *et al.* (2009) Phylogenetic diversity and metabolic potential revealed in a glacier ice metagenome. *Appl. Environ. Microbiol.*, **75**, 7519-7526.

Singleton,C.M. *et al.* (2021) Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing. *Nat. Commun.*, **12**, 2009.

Soares,S.C. *et al.* (2013) The Pan-Genome of the Animal Pathogen Corynebacterium pseudotuberculosis Reveals Differences in Genome Plasticity between the Biovar ovis and equi Strains. *PLoS One*, **8**, e53818.

Spring-Pearson,S.M. *et al.* (2015) Pangenome analysis of burkholderia pseudomallei: Genome evolution preserves gene order despite high recombination rates. *PLoS One*, **10**, e0140274.

Stamatakis,A. (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312-1313.

Surujonu,D. *et al.* (2020) Boundary-Forest Clustering: Large-Scale Consensus Clustering of Biological Sequences. *bioRxiv*. doi: https://doi.org/10.1101/2020.04.28.065870

Taylor-Brown,A. *et al.* (2017) Culture-independent metagenomics supports discovery of uncultivable bacteria within the genus Chlamydia. *Sci. Rep.*, **7**, 10661.

Tett,A. *et al.* (2019) The Prevotella copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations. *Cell Host Microbe*, **26**, 666-679.

Tettelin,H. *et al.* (2008) Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.*, **11**, 472-477.

Tettelin,H. *et al.* (2005) Genome analysis of multiple pathogenic isolates of

Streptococcus agalactiae: Implications for the microbial 'pan-genome'. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 13950-13955.

Tonkin-Hill,G. *et al.* (2020) Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.*, **21**, 180.

de Toro,M. *et al.* (2014) Plasmid Diversity and Adaptation Analyzed by Massive Sequencing of Escherichia coli Plasmids. *Microbiol. Spectr.*, **2**, PLAS–0031.

Touchon,M. *et al.* (2009) Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLoS Genet.*, **5**, e1000344.

Trainor,E.A. *et al.* (2015) Bordetella pertussis transmission. *Pathog. Dis.*, **73**, ftv068.

Tully,B.J. *et al.* (2018) The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci. Data*, **5**, 170203.

Tyson,G.W. *et al.* (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, **428**, 37-43.

Utter,D.R. *et al.* (2020) Metapangenomics of the oral microbiome provides insights into habitat adaptation and cultivar diversity. *Genome Biol.*, **21**, 293.

Vavourakis,C.D. *et al.* (2018) A metagenomics roadmap to the uncultured genome diversity in hypersaline soda lake sediments. *Microbiome*, **6**, 168.

Venter,J.C. *et al.* (2004) Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science*, **304**, 66-74.

Vernikos,G. *et al.* (2015) Ten years of pan-genome analyses. *Curr. Opin. Microbiol.*, **23**, 148-154.

Vernikos,G.S. (2020) A review of pangenome tools and recent studies. In, *The Pangenome: Diversity, Dynamics and Evolution of Genomes*, pp 89-112.

Vieira,G. *et al.* (2011) Core and panmetabolism in Escherichia coli. *J. Bacteriol.*, **193**, 1461-1472.

Virtanen,P. *et al.* (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods*, **17**, 261-272.

van Vliet,A.H.M. (2017) Use of pan-genome analysis for the identification of lineage-specific genes of Helicobacter pylori. *FEMS Microbiol. Lett.*, **364**, fnw296.

Van Der Walt,S. *et al.* (2011) The NumPy array: A structure for efficient numerical computation. *Comput. Sci. Eng.*, **13**, 22-30.

Wang,Y. *et al.* (2019) Genomics insights into ecotype formation of ammonia-oxidizing archaea in the deep ocean. *Environ. Microbiol.*, **21**, 716-729.

Watson,M. (2021) New insights from 33,813 publicly available metagenome-assembled-genomes (MAGs) assembled from the rumen microbiome. *bioRxiv*. doi: https://doi.org/10.1101/2021.04.02.438222.

Weigand,M.R. *et al.* (2017) The history of Bordetella pertussis genome evolution includes structural rearrangement. *J. Bacteriol.*, **199**, e00806-16.

Wickham,H. (2016) ggplot2 Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4.

Wilkins,L.G.E. *et al.* (2019) Metagenome-assembled genomes provide new insight into the microbial diversity of two thermal pools in Kamchatka, Russia. *Sci. Rep.*, **9**, 3059.

Wu,H. *et al.* (2021) Toward a high-quality pan-genome landscape of Bacillus subtilis by removal of confounding strains. *Brief. Bioinform.*, **22**, 1951-1971.

Wu,Y.W. *et al.* (2014) MaxBin: An automated binning method to recover individual

genomes from metagenomes using an expectation-maximization algorithm. *Microbiome,* **2**, 26.

Zhang,Y. and Sievert,S.M. (2014) Pan-genome analyses identify lineage- and niche-specific markers of evolution and adaptation in Epsilonproteobacteria. *Front. Microbiol.*, **5**, 110.

Zhao,Y. *et al.* (2012) PGAP: Pan-genomes analysis pipeline. *Bioinformatics*, **28**, 416-418.

Zhou,Z. *et al.* (2020) Accurate reconstruction of bacterial pan- and core genomes with PEPPAN. *Genome Res.*, **30**, 1667-1679.

Zhou,Z. *et al.* (2018) Pan-genome Analysis of Ancient and Modern Salmonella enterica Demonstrates Genomic Stability of the Invasive Para C Lineage for Millennia. *Curr. Biol.*, **28**, 2420-2428.

APPENDICES

APPENDIX A

PUBLICATION SUMMARY

Appendix A. Summary of 40 papers that using pan-genome analysis on MAGs.

PMID/DOI	Title	Authors	Year	Journal	Source of Metagenomes	Tools used in pan-genome	Parameters	Downstream analysis
33323129	Metapangenomics of the oral microbiome provides insights into habitat adaptation and cultivar diversity	Utter, et al.	2020	Genome Biology	human tongue and dental plaque	Anvi'o	MCL inflation value of 2	Functional enrichment analysis (Pfam)
33323122	Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome	Shaiber, et al.	2020	Genome Biology	human tongue and dental plaque	Anvi'o	MCL inflation value of 2, minimum gene occurrence of 2	Functional enrichment analysis (COG)
33295091	Metagenomic insights into the metabolism and evolution of a new *Thermoplasmata* order (*Candidatus Gimiplasmatales*)	Hu, et al.	2020	Environmental Microbiology	black-odorous aquatic river	GET_HOMOLOGUES	Diamond: e-value 1e-05, a minimum of 75% coverage in BLAST pairwise alignments	\
33239396	Deep metagenomics examines the oral microbiome during dental caries, revealing novel taxa and co-occurrences with host molecules	Baker, et al.	2021	Genome Research	human saliva	Anvi'o	eggnog-mapper for annotation	Functional enrichment analysis (COG)

33127895	Gene duplication drives genome expansion in a major lineage of Thaumarchaeota	Sheridan, et al.	2020	Nature Communications	river	Roary	50% clustering identity; MCL inflation value of 1.5, 85% core gene threshold	Phylogenetic analysis
33067437	Large scale genome reconstructions illuminate Wolbachia evolution	Scholz, et al.	2020	Nature Communications	different host speceis like Drosophila (Downloaded from NCBI)	Roary	80% clustering identity and no paralog splitting. The core genes present in at least one representative genomes in each *Wolbachia* supergroups.	Functional enrichment analysis, Phylogenetic analysis.
33065016	Newly Explored Faecalibacterium Diversity Is Connected to Age, Lifestyle, Geography, and Disease	De Filippis, et al.	2020	Current Biology	human and animal guts	Roary	95% clustering identity, 95% core gene threshold	Phylogenetic analysis (Species-specific marker genes from core genes)
32934112	Defining Genomic and Predicted Metabolic Features of the Acetobacterium Genus	Ross, et al.	2020	mSystems	wastewater and bioreactor	BPGA	clustering cutoffs: 10%-99%. 100% core gene threshold	Functional prediction (KEGG)
32690973	A unified catalog of 204,938 reference genomes from the	Almeida, et al.	2020	Nature Biotechnology	human gut	Roary	Prokka: -c -m -p single.	Functional enrichment analysis (COG)

	human gut microbiome						Roary: 95% clustering identity; 90% core gene threshold; Do not split paralogs	
32636492	Alternative strategies of nutrient acquisition and energy conservation map to the biogeography of marine ammonia-oxidizing archaea	Qin, et al.	2020	The ISME Journal	activated sludge, water, hot springs, etc.	OrthoMCL	50% identity and 50% coverage for clustering; 100% core gene threshold	Phylogenomic analysis (Phylogenetic tree built based on conserved single-copy homologous protein)
32631866	Genomic Characteristics of a Novel Species of Ammonia-Oxidizing Archaea from the Jiulong River Estuary	Zou, et al.	2020	Applied and Environmental Microbiology	water and sediment	Anvi'o	\	\
32348781	A Genomic Toolkit for the Mechanistic Dissection ofIntractable Human Gut Bacteria	Bisanz, et al.	2020	Cell Host & Microbe	human gut	ProteinOrtho6	60% identity and 80% coverage for clustering; 100% core gene threshold	Functional enrichment analysis (KEGG)
32316533	Mechanisms Underlying the Rhizosphere-To-Rhizoplane	Zhang, et al.	2020	Microorganisms	citrus trees	GET_HOMOLOGUES	50% identity and 50% coverage for clustering	Phylogenetic analysis (single-copy core genes)

	Enrichment of Cellvibrio Unveiled by Genome-Centric Metagenomics and Metatranscriptomics							
32291353	Selection Is a Significant Driver of Gene Gain and Loss in the Pangenome of the Bacterial Genus Sulfurovum in Geographically Distinct Deep-Sea Hydrothermal Vents	Moulana, et al.	2020	mSystems	diffuse flow hydrothermal vent fluids	Anvi'o	Diamond: E value 1e-05; Anvi'o: minbit value of 0.5 and MCL inflation value of 2	Functional enrichment analysis (COG)
32169939	Temperature and Nutrient Levels Correspond with Lineage-Specific Microdiversification in the Ubiquitous and Abundant Freshwater Genus Limnohabitans	Props, et al.	2020	Applied and Environmental Microbiology	freshwater	Anvi'o	minbit value of 0.5 and MCL inflation value of 2	Functional enrichment analysis (KEGG)
32103005	A comprehensive non-redundant gene catalog reveals extensive within-community intraspecies diversity in the human vagina.	Ma, et al.	2020	Nature Communications	human vagina	CD-HIT-EST	clustering: >=99 bp long, sequence identity >=95% , >=90% of the shorter gene length.	\
31757822	Comparative Genomics Guides Elucidation of	Kirmiz, et al.	2020	Applied and Environment	children gut	Anvi'o	minbit value of 0.5, MCL inflation value	Functional prediction (COG).

	Vitamin B12 Biosynthesis in Novel Human-Associated Akkermansia Strains			al Microbiology			of 10, and use-ncbi-blast. 100% core gene threshold.	Phylogenetic analysis.
31607556	The Prevotella copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations	Tett, et al.	2019	Cell Host & Microbe	human gut	Roary	90% clustering identity; 90% core gene threshold.	Functional prediction (using EggNOG).
31118472	Uncultured Nitrospina-like species are major nitrite oxidizing bacteria in oxygen minimum zones	Sun, et al.	2019	The ISME Journal	Seawater	Anvi'o	BLASTp: e-value 10^{-10} Anvi'o: MCL inflation value of 10.	Phylogenetic analysis
30867587	New insights from uncultivated genomes of the global human gut microbiome	Nayfach, et al.	2019	Nature	human gut	VSEARCH	90% DNA identity and 50% alignment cut-offs	Phylogenetic analysis
30837458	The Wolbachia mobilome in Culex pipiens includes a putative plasmid	Reveillaud, et al.	2019	Nature Communications	ovary of wild Culex pipiens mosquitoes	Anvi'o	minbit value of 0.5, MCL inflation value of 10, and use-ncbi-blast. 100% core gene threshold.	Functional enrichment analysis (COG)
30816235	Metagenome-assembled genomes provide new insight into the microbial diversity of two	Wilkins, et al.	2019	Scientific Reports	hydrothermal pools	Anvi'o	MCL inflation value of 10	Functional enrichment analysis (COG and KEGG)

	thermal pools in Kamchatka, Russia							
30768760	Deep-sea hydrothermal vent metagenome-assembled genomes provide insight into the phylum Nanoarchaeota	St John, et al.	2019	Environmental Microbiology Reports	deep-sea hydrothermal vent	Anvi'o	\	Functional prediction (COG and Pfam)
30661755	Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle	Pasolli, et al.	2019	Cell	human gut, oral cavity, skin and vagina	Roary	Prokka (default); Roary: 95% clustering identity; 99% core gene threshold	phylogenetic analysis; functional profiles (KEGG functions)
30592124	Genomics insights into ecotype formation of ammonia-oxidizing archaea in the deep ocean	Wang, et al.	2019	Environmental Microbiology	ocean	Anvi'o	BLASTP against the COG database.	Functional enrichment analysis (COG and KEGG)
30448738	Genome-resolved metagenomic analysis reveals roles of microbial community members in full-scale seawater reverse osmosis plant	Rehman, et al.	2018	Water Research	raw seawater, fouled reverse osmosis membranes and brine reject water.	Roary	\	Functional prediction.

30394652	Comparative genomics and physiology of the genus Methanohalophilus, a prevalent methanogen in hydraulically fractured shale	Borton, et al.	2018	Environmental Microbiology	hydraulically fractured shale	Anvi'o	MCL inflation value of 10; 100% core gene threshold	Phylogenetic analysis (single copy core genes and concatenated ribosomal amino acid sequences)
30319574	To B or Not to B: Comparative Genomics Suggests Arsenophonus as a Source of B Vitamins in Whiteflies.	Santos-Garcia, et al.	2018	Frontiers in Microbiology	whitefly	OrthoMCL	100% core gene threshold	Functional enrichment analysis (COG)
30117250	Ecological and genomic features of two widespread freshwater picocyanobacteria	Cabello-Yeves, et al.	2018	Environmental Microbiology	(Downloaded from NCBI)	GET_HOMOLOGUES	100% core gene threshold	
30083144	New Biological Insights Into How Deforestation in Amazonia Affects Soil Microbial Communities Using Metagenomics and Metagenome-Assembled Genomes	Kroeger, et al.	2018	Frontiers in Microbiology	soil	Anvi'o	Used the NCBI-blast; 66% core gene threshold.	Phylogenetic analysis
30033331	Pan-genome Analysis of Ancient and Modern Salmonella enterica	Zhou, et al.	2018	Current Biology	human teeth and long bones of skeletons	BLASTN	At least 50% sequence length coverage,	\

	Demonstrates Genomic Stability of the Invasive Para C Lineage for Millennia						70% nucleotide identity.	
29066755	Genomic variation in microbial populations inhabiting the marine subseafloor at deep-sea hydrothermal vents	Anderson, et al.	2017	Nature Communications	hydrothermal vents	ITEP	MCL inflation value of 2, a maxbit score of 0.3.	Functional enrichment analysis (COG)
28458657	Comparative Metagenomics of the Polymicrobial Black Band Disease of Corals	Meyer, et al.	2017	Frontiers in Microbiology	coral surface	Anvi'o; Roary	Anvi'o: MCL inflation value of 2, and use NCBI blastp; Core genes present in at least 5/7 of the MAGs.	Functional enrichment analysis (KEGG)
DOI:https://doi.org/10.1101/2020.08.09.243345	Diversity and biogeography of Woesearchaeota: A comprehensive analysis of multi-environment data	Xian, et al.	2020	\	sea water, rhizosphere and sediment	BPGA	USEARCH: orthologous clustering with 50% sequence identity. 100% core gene threshold	Functional enrichment analysis (COG and KEGG)
DOI: https://doi.org/10.21203/rs.3.rs-60068/v1	Long-read metagenomics retrieve complete single-contig bacterial genomes from canine feces	Cusco, et al.	2020	\	canine fecal samples; animal gut	Anvi'o	MCL inflation value of 10	Functional predictions (COG)

DOI: https://doi.org/10.1101/2021.03.02.433653	Adaptive ecological processes and metabolic independence drive microbial colonization and resilience in the human gut	Watson, et al.	2021	\	human gut	Anvi'o	Minimum gene occurrence of 2.	Phylogenetic analysis, functional enrichment analysis (KEGG)
DOI: https://doi.org/10.1101/2020.12.11.421487	The hidden pangenome: comparative genomics reveals pervasive diversity in symbiotic and free-living sulfur-oxidizing bacteria	Ansorge, et al.	2020	\	hydrothermal vents	BPGA	USEARCH: orthologous clustering with 50% sequence identity. 100% core gene threshold	Functional enrichment analysis (COG and KEGG)
DOI: https://doi.org/10.1101/2020.07.02.185041	Metagenome-assembled genomes from Monte Cristo Cave (Diamantina, Brazil) reveal prokaryotic lineages as functional models for life on Mars	Bendia, et al.	2020	\	cave	Anvi'o	100% core gene threshold	Functional enrichment analysis (KEGG)
DOI: https://doi.org/10.1101/2021.01.19.427344	Large lakes harbor streamlined free-living nitrifiers	Podowski, et al.	2021	\	water samples from lakes	Anvi'o	minbit value of 0.5, MCL inflation value of 2, and minimum gene occurrence of 1.	Functional predictions (KEGG)

DOI:https://doi.org/10.1016/j.margen.2019.04.010	Reconstruction and in silico analysis of new Marinobacter adhaerens t76_800 T with potential for long-chain hydrocarbon bioremediation associated with marine environmental lipases	Lopes, et al.	2020	Marine Genomics	marine	Anvi'o; BPGA	100% core gene threshold	Functional prediction (COG and dbCan)

APPENDIX B
CASE STUDY FOR CORE GENE LOSS

Appendix B. Case study – the loss core gene in a contaminated *E. coli* genome (GCF_900149915.1_ 128_cut_98.97comp_3.06cont). Blue frames represent input files. Green frames represent outputs. Orange frames represent the example for core gene loss in an E. coli genome (numbers for gene clusters or reasons for core gene loss)