UNIVERSITY OF CALIFORNIA SAN DIEGO

Pangenome Analysis of H. pylori: A Systematic Approach to Study Genetic Variability, Phylogenetic Grouping, and Regulatory Networks

A Thesis submitted in partial satisfaction of the requirements
for the degree Master of Science

in

Bioengineering

by

Qiangsheng Yin

Committee in charge:

Professor Bernhard Palsson, Chair
Professor Ludmil B. Alexandrov
Professor Prashant Mali

2024

The Thesis of Qiangsheng Yin is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to everyone who contributed to the success of this research. First and foremost, I am deeply appreciative of my supervisor and my mentors, Dr. Bernhard Palsson, Annie Yuan, and Siddhardth Chauhan, whose guidance and expertise have been invaluable throughout this study. Their suggestions greatly enhanced the methodology of this research, particularly in the application of Non-negative Matrix Factorization (NMF) and iModulon analysis.

I would love to extend my gratitude to my parents and friends for their encouragement during the research and writing process. Your unwavering support provided me with the strength and motivation needed to complete this work.

ABSTRACT OF THE THESIS

Pangenome Analysis of H. pylori: A Systematic Approach to Study Genetic Variability, Phylogenetic Grouping, and Regulatory Networks

by

Qiangsheng Yin

Master of Science in Bioengineering

University of California San Diego, 2024

Professor Bernhard Palsson, Chair

We proposed a systematic method to understand *H. pylori*'s genetic diversity, phylogenetic clustering, and regulatory relations. This study provides a pangenome analysis of over 1,300 complete *H. pylori* strains, which is over ten times higher than previous studies, significantly expanding the scope of genetic exploration. We identified 1,015 core genes, 986 accessory genes, and 38,357 rare genes. Non-negative matrix factorization (NMF) was used for phylogenetic clustering, allowing us to decompose the accessory gene matrix for a better

mathematical representation. We applied a Random Forest Classifier to characterize the genetic basis of these phylogroups, highlighting the genes that contribute most significantly to phylon differentiation. Finally, by integrating pangenome data with RNA-seq analysis, we created a multi-strain dataset with enhanced statistical power and comparability to better understand gene functionality and discover new regulatory networks and to address the challenge of limited availability of single-strain transcriptomic data in many bacterial species. This approach creates a comprehensive framework for H. pylori studies using public genomic and transcriptomic data, offering a scalable model for similar studies in other bacterial species.

Chapter 1 Introduction

H. pylori: An overview

Add Helicobacter pylori (H. pylori) is a gram-negative, spiral-shaped bacterium that colonizes the human stomach. H. pylori is a widespread bacterial pathogen that affects approximately 50% of the world's population, with a higher prevalence in developing countries[1,2]. This bacterium is a significant cause of chronic gastritis and is strongly associated with the development of peptic ulcer disease and gastric cancer[3,4]. Recent studies have expanded the clinical significance of H. pylori, linking it to extragastric conditions such as iron deficiency anemia and idiopathic thrombocytopenic purpura, highlighting its diverse impact on human health [5]. The transmission of H. pylori predominantly occurs during childhood, especially in low socioeconomic conditions where poor hygiene is prevalent[6]. Despite some declines in infection rates in industrialized countries, H. pylori remains a significant global health concern[7].

H. pylori has numerous mechanisms that enable it to survive in the acidic gastric environment. One of its key survival strategies is the production of urease, an enzyme that catalyzes the conversion of urea to ammonia and carbon dioxide. This reaction neutralizes gastric acid, creating a more favorable environment for bacterial colonization[8]. Biofilm formation is another critical adaptation that protects the bacteria from the acidic gastric environment, immune system attacks, and antibiotic treatments[9]. The ability to form biofilms and increasing antibiotic resistance poses a challenge to the effective eradication of H. pylori[10].

The bacterium can also alter its gene expression in response to acidic conditions and manipulate host immune responses to persist long-term, further aiding its survival and pathogenicity[11,12]. These adaptations enable H. pylori to colonize the stomach for decades, potentially leading to severe gastric diseases[13].

H. pylori genome

H. pylori is a bacterium with a highly variable but relatively small genome, ranging from 1.5 to 1.7 million base pairs[14]. Despite its small size, it contains unique regions associated with pathogenicity, such as the cagA pathogenicity island, which is a Type VII secretion system. This genetic variability significantly contributes to the bacterium's adaptability and pathogenic potential.

Despite recent advances in genomic sequencing and analysis, a significant portion of H. pylori genes remains functionally uncharacterized. More than one-third of the genes in our study are poorly annotated, indicating the need for further research to uncover detailed gene functions, virulence factors, and metabolic pathways to develop effective treatments and managing gastroduodenal disorders associated with H. pylori infection.

Pangenome studies analyze the complete set of genes within a species, encompassing both core genes present in all individuals and accessory genes found in some[15]. This approach has revolutionized microbial research, providing insights into the genetic diversity, evolution, and functional potential of bacterial species[16,17]. Pangenome analysis has been widely used to redefine pathogenic species and examine genomic diversity in organisms such as Escherichia coli[18], highlighting its broad applicability across plants and animals[19].

The pangenome analysis of H. pylori offers valuable insights into its genetic diversity, pathogenicity, and evolutionary history. Studies have shown that the core genome of H. pylori consists of conserved gene families that make up a substantial portion of the bacterium's genetic content[20,21]. These core genes are essential for basic cellular functions and survival in the gastric environment. The accessory genome of H. pylori, which varies among strains, contributes to the

bacterium's adaptability. Thus, characterizing the core and accessory genome of H. pylori helps identify differences among strains and discover potential drug targets.

Current State of H. pylori Pangenome Studies

The study by Amjad Ali et al. selected 39 complete Helicobacter pylori genomes from the NCBI database to conduct pangenome analysis[20]. The researchers predicted open reading frames (ORFs) from the selected DNA sequences. Phylogenetic analysis was conducted using 16S rRNA gene sequences to establish evolutionary relationships among the strains. Comparative analysis of proteomes was performed using BLAST, and the core genome and pan-genome were identified using a 50/50 clustering approach. Amjad Ali et al. identified the core genome of H. pylori, consisting of 1,193 genes. This core genome accounts for 77% of the average genome and 45% of the global gene repertoire of H. pylori. While the study provided valuable insights into the core genome and potential therapeutic targets, it was limited by the relatively small number of genomes analyzed. Moreover, the study primarily focused on the core genome, leaving the accessory genome relatively unexplored.

Another study led by van Vliet expanded the scope of pangenome analysis by analyzing 346 high-quality H. pylori genomes[21]. Phylogenetic clustering was performed using core genome SNPs and whole-genome purine/pyrimidine (RY) words. Genome annotation and pan-genome analysis were conducted using Roary. The study revealed that lineage-specific genes can contribute to variations in virulence and disease outcomes. For example, the FecA2, ferric citrate receptor gene, was absent in hspAmerind genomes but present in all other lineages, suggesting potential differences in iron acquisition among populations. This study highlighted the genetic diversity of H. pylori in different geographic locations. However, the high level of allelic

variability in H. pylori genomes limits the completeness of the analysis. The pangenome analysis may overestimate differences between H. pylori genes due to the high levels of sequence variation, as genes with similar functionality may be classified as distinct.

Cao conducted another comprehensive analysis of 99 Helicobacter genomes, including 75 H. pylori and 24 non-pylori Helicobacter species (NPHS) genomes[22]. This study aimed to explore the genomic diversity and adaptability of H. pylori. The researchers used Glimmer version 3.02 to predict open reading frames (ORFs) and constructed a phylogenetic tree using 16S rRNA genes, with Campylobacter species as outgroups. Orthologous group analysis was performed using OrthoMCL to identify core and accessory genomes. The study revealed that H. pylori has an open and diverse genome with 1,173 conserved protein families (core genomes). The lack of functional characterization for many genes also poses a problem in this study, suggesting further research is needed to understand H. pylori pathogenicity and adaptability better.

In another study, Uchiyama analyzed the pangenome of 30 completely sequenced H. pylori strains belonging to various phylogeographic groups and identified 991 accessory orthologous groups (OGs) that were not fully conserved[23]. A novel method was developed to evaluate the mobility of genes, using the gene order in syntonically conserved regions to classify genes into five classes: core, stable, intermediate, mobile, and unique. The study found that phylogenetic networks based on the gene content of core and stable classes were highly congruent with those created from fully conserved core genes. In contrast, the Intermediate and Mobile classes showed different topologies. The generality and usability of methods used in gene class assignment still need to improve to apply to other bacterial species.

Current Study and Its Contributions

My study addresses some of the current limitations of pangenome analysis on H. pylori and other bacterial species. By using a gene clustering method, CD-HIT, to handle the large DNA-seq dataset, we constructed a pangenome from over 1,300 H. pylori complete strains. This is ten times more than previous studies, offering a more comprehensive identification of the species' core, accessory, and rare genomes, which can better help us understand the genetic variability of H. pylori.

The innovative application of non-negative matrix factorization (NMF) to the accessory genome matrix was a pivotal aspect of our study. NMF provided a robust framework for identifying phylogenetic groups based on shared gene content, which is especially relevant given the high genetic variability among H. pylori strains. By decomposing the matrix into components, we could characterize the genetic basis of these groups and gain insights into how different strains are related at a genomic level.

We applied two methods for the genetic basis characterization. First, the top-down approach in gene identification was beneficial in highlighting specific gene functions related to particular phylogroups. This method is particularly effective for well-annotated genomes where functional categories are well-defined. However, given the limitations in annotation for many H. pylori genes, the complementary use of a Random Forest Classifier proved helpful. This machine learning approach allowed for an unbiased identification of genes with high feature importance, thus highlighting genes that contribute significantly to the differentiation of phylogroups.

The integration of RNA-seq data using iModulon analysis added another layer of understanding to our study. By analyzing gene expression patterns across multiple strains, we

could identify regulatory networks that are conserved across strains. This is crucial for understanding how H. pylori regulates gene expression in response to different environmental changes. The use of iModulon analysis on combining core and accessory genomes significantly improves the explained variance, offering a clearer insight into the transcriptional network of H. pylori.

Overall, this study addresses some gaps in previous research by analyzing a larger number of strains and employing advanced computational techniques such as NMF, Random Forest Classifier, and iModulon analysis to explore the gene diversity and functionality of H. pylori. These methods provide a systematic approach to analyze bacterium's pangenome.

Chapter 2 Methods

Data Acquisition and Filtration

The H. pylori sequences were downloaded from the Bacterial and Viral Bioinformatics Resource Center (BV-BRC)[24] and the National Center for Biotechnology Information (NCBI) databases[25,26]. The H. pylori strains downloaded covered different clinical backgrounds and geographic regions, guaranteeing a representative and diverse dataset for our investigation.

Several filtering criteria were applied to ensure the quality of the sequence data. First, sequences were filtered based on L50/N50 metrics, which assess the quality of genome sequencing by evaluating the counts of contigs that cover 50% of the genome. The N50 value of a reference genome was used as a benchmark, and strains with N50 values smaller than 0.85 times that reference value were removed.

Next, the CheckM completeness and contig count distributions were inspected to remove strains with outlier values, excluding genomes with significant gaps. Highly fragmented sequences were also excluded by setting a maximum ceiling for the number of contigs and removing strains with contig numbers larger than two times the median contig count in the dataset.

Genomes with GC content outside the range of 35% to 40% were filtered out, as these could indicate sample contamination or sequencing errors. After downloading the genome data, additional quality control measures were conducted to remove sequences with abnormally short sequences, duplicated samples, or other quality issues to prevent inaccuracies in the analysis. In the last step of filtering, a manual inspection was performed to examine metadata for inconsistencies or anomalies that automated filters might have missed, including verifying strain metadata against established sources.

Mash Clustering and Filtration

We implemented Mash clustering to effectively group the H. pylori strains in our study for the reference number of clusters and initial screening. Mash is an alignment-free method based on the MinHash algorithm, which condenses large genomic sequences into small, representative sketches for quick estimation of pairwise distances[27]. This approach is particularly advantageous for large-scale analyses, offering comparable accuracy to alignment-based methods but with significantly faster computation times.

First, all complete sequences from the filtered dataset were used to generate a pairwise distance matrix using Mash. To filter the strains based on these Mash distances, a low threshold was set at 0.05, and a high threshold was set at the 97th percentile of Mash distances relative to the reference genome. The 97th percentile threshold was determined based on the Mash distance distribution to avoid the second peak and to ensure the consistency of our dataset.

The Mash distance values were then converted into a Pearson correlation distance matrix, and the genomes were clustered using the hierarchical clustering function of the SciPy package in Python with Ward's linkage method[28]. Initial clusters were evaluated, and small clusters containing fewer than five strains were filtered out. The Pearson correlation matrix was regenerated, and clustering was performed again. This iterative process continued until robust clusters emerged, indicated by a stable number of clusters and the absence of small clusters. The number of clusters obtained from this iterative Mash clustering process was then used as a reference for subsequent non-negative matrix factorization (NMF) analysis.

Pangenome Construction

To construct the pangenome of more than one thousand H. pylori strains, we used the CD-HIT algorithm, a widely used program for clustering DNA sequences, to reduce redundancy and improve the efficiency of sequence analyses[29]. CD-HIT groups sequences according to a specified identity threshold, generating representative sequences for each cluster. This approach significantly reduced the number of genes in our pangenome, allowing us to consider only the representative genes and construct the pangenome matrix further.

Several identity thresholds from 70% to 90% were evaluated to determine the best fit. Metrics such as the number of gene clusters identified were examined, and an 80% identity threshold was chosen, the same as in previous studies. This threshold ensures that only highly similar sequences are clustered together, reducing dataset complexity while preserving meaningful biological variation.

The representative genes were then classified into core, accessory, and rare genomes. Core genes are typically present in nearly all strains and are considered essential for H. pylori survival. Rare genes are limited to one or a few strains, highlighting unique genetic elements that could be important for specific niche adaptations or recent acquisitions. The remaining genes are accessory genes found in particular groups but not all strains. These accessory genes are the focus of our study to classify different phylogroups, as the accessory genome does not contain either highly conserved or highly variant genes.

Classification boundaries for core, accessory, and rare genomes were determined by examining the gene frequency distribution—the number of genomes in which each gene is found. This distribution typically shows two peaks: one for rare genes and one for highly

conserved core genes. Thus, the cumulative gene frequency distribution forms an inverse sigmoidal curve[30]. Three phases in the curve represent three gene categories: the initial log phase for rare genes, the middle stationary phase for accessory genes, and the second log phase for core genes.

The cumulative gene frequency distribution was fitted to the sum of two power functions to capture these two peaks. The cutoff boundaries were defined by the relative distance to the inflection point in the simulated curve. Genes with frequencies less than 90% from the inflection point to the minimum frequency were classified as rare genes. Genes with higher frequencies than 90% from the inflection point to the maximum frequency were classified as core genes, and genes in between were accessory genes. Additionally, Heaps' law was applied to test the openness of the pangenome, helping us understand whether the pangenome is finite or continues to expand as more genomes are added.

The results from CD-HIT were transformed into a presence/absence binary gene matrix, with each column representing a gene cluster and each row representing a strain. The accessory gene matrix was prioritized for further analysis, as the core gene matrix contains only highly conserved genes, while the rare gene matrix is sparse and predominantly filled with zeroes.

Gene annotation and enrichment analysis

To ensure consistency during analysis, Prokka was used to annotate all genomes. Prokka is a rapid prokaryotic genome annotation tool that assigns functions to genes based on homology searches against several databases, including UniProt and Pfam[31]. All representative genes generated by CD-HIT were further annotated using eggNOG mapper, which assigns COG

(Clusters of Orthologous Groups) categories to each gene, enhancing the understanding of each gene's role within the genome[32]. This comprehensive annotation process provided additional information to aid in later cluster characterization and genetic basis analysis.

Following COG annotation, an enrichment analysis was conducted to identify overrepresented and underrepresented genes within the core and accessory genomes. Fisher's exact test was applied to compare the gene frequency of each COG category in the core and accessory genomes against a background frequency derived from the entire pangenome. The p-values obtained from Fisher's exact test were adjusted for multiple tests using the family-wise error rate (FWER). To quantify the extent of overrepresentation, the log2 odds ratio (LOR) was calculated for each COG category. This analysis helped identify functional differences between the core and accessory genomes, providing insights into the roles these genes may play in H. pylori's adaptability, pathogenicity, and survival.

Non-negative matrix factorization (NMF)

Non-negative matrix factorization (NMF) is a technique for decomposing a matrix into two lower-dimensional, non-negative matrices. This method is particularly effective for uncovering patterns and extracting meaningful components from complex datasets, such as genomic data. NMF is especially advantageous for binary matrices as it maintains the non-negative nature of the data, making it ideal for analyzing gene presence-absence matrices in pangenome studies.

We applied NMF to the accessory gene matrix P, which consists of binary values indicating the presence or absence of genes across different strains[33]. Using the scikit-learn

implementation of NMF, the decomposition was performed 50 times at ranks around the suggested number of Mash clusters. Each run decomposed the original matrix P into two non-negative matrices, L and A, where $P = LA$[33]. The L matrix has the dimensions (number of genes) × (rank), and A has the dimensions (rank) × (number of strains). The rank represents the number of computed phylogroups. The L matrix includes gene weightings for each computed phylon, while the A matrix represents the strain-specific affinity to a phylogroup. In biological terms, gene weightings in the L matrix define how much genes contribute to the specific phylon classification, revealing the common gene sets in certain phylogroups. The values in the A matrix indicate how closely strains belong to the defined phylogroups.

To determine the optimal rank for NMF decomposition, various ranks were evaluated, and the best one was selected based on the performance of the L and A matrices in reconstructing the P matrix using the F1-score as a metric. The best run at each rank was chosen for normalization by evaluating metrics such as the Frobenius norm, sum of squared residuals, and root-mean-square error. The L matrix was normalized by dividing its values by the 99th percentile for each column, and the A matrix was scaled correspondingly by row using the same normalization factors to ensure that most values were between 0 and 1.

After normalization, the L and A matrices were binarized using k-means clustering with k=3 clusters. Genes were divided into three clusters for each column in the L matrix using the scikit-learn k-means package[34]. The genes in the cluster with the highest average mean were assigned a value of 1, while those in the other two clusters were assigned a value of 0. This binarization method was applied to the A matrix by row. The zeroes and ones in the binarized L and A matrices can be simply interpreted as whether phylogroups contain a strain and whether genes are included in a phylogroup.

Characterization

Following the NMF analysis, the phylon and their associated genes were characterized to understand the biological significance of the identified phylogroups. This process involved several approaches to understand how specific genes contribute to phylon classification.

The characterization of the binarized A matrix was based on metadata, such as geographic location for H. pylori strains. By analyzing the strains within each phylon, commonalities and patterns were identified, providing insights into how geographic and other metadata correlate with the phylogenetic groupings.

Next, the L matrix was examined to determine the genetic basis of the identified phylogroups and answer the question of what genes distinguish these phylogroups. Genes with the highest variances in the L matrix were identified and examined for enrichment in any COG categories. This analysis helped understand which genes were most variable across phylogroups and identify which sets of genes require further investigation in later analyses.

The binarized L matrix was hierarchically clustered to reveal phylogenetic relationships and a top-down approach was employed to identify exclusive genes in each phylon or phylogroup from the same node. A BLAST search was conducted to identify common outer membrane proteins and their orthologs, providing a profile of these highly variable genes. By examining the presence-absence patterns of genes within each phylon, it was possible to determine which genes were unique to or shared among different groups, shedding light on the genetic basis of phylogroup-specific traits.

In cases where there is no well-annotated database or specific enriched COG categories to explore, a more systematic approach was developed to identify phylon-specific genes. A random forest classifier was applied to the original L matrix to find genes with the highest feature importance. Random forest is an ensemble learning method that constructs multiple decision trees during training and merges their predictions to improve model performance. This method was chosen over linear SVM and single decision trees due to its ability to handle non-linear relationships and reduce overfitting.

To use the random forest classifier, customized binary labels were created where the subset of interest was labeled as 1 and all other samples as 0, and the classifier was run 100 times. In each iteration, the classifier was fitted on the filtered dataset, and feature importances were accumulated over all runs. The accumulated feature importances were averaged to obtain the final importance score for each gene, and the top features contributing most to distinguishing the subset from the rest were identified. The standard deviation of feature importance across runs was calculated to assess their stability, ensuring robust and reliable identification of important features.

Multi-strain dataset construction and iModulon analysis

To explore the regulatory networks and gene functionality of Helicobacter pylori, we constructed a multi-strain RNA-seq dataset using a pangenome approach. The first step in constructing the multi-strain RNA-seq dataset involved identifying core and accessory genes across the chosen strains. We conducted a bi-directional BLAST search to identify homologous genes that allowed us to create a unified gene set for expression profiling. The RNA-seq data from the selected strains were then integrated. By excluding rare genes, we minimized noise and

improved the statistical power, enabling the detection of subtle gene expression variations that might be overlooked in single-strain analyses.

Once the dataset was compiled, we used iModulon analysis to uncover independently modulated gene sets, known as iModulons. This method, based on independent component analysis (ICA), allows for the identification of co-regulated genes that form distinct regulatory modules, providing insights into H. pylori's transcriptional networks[35]. By comparing the activities of these iModulons across different experimental conditions and strains, we gained a deeper understanding of the regulatory roles of specific genes and pathways, shedding light on the complex mechanisms governing H. pylori's adaptability and pathogenicity.

Overall, this comprehensive multi-strain RNA-seq dataset and subsequent iModulon analysis offer a novel perspective on the transcriptional landscape of H. pylori, providing a foundation for future research into its genetic and functional complexity. This methodology paves the way for similar approaches in other bacterial species, particularly those with limited available RNA-seq data.

Chapter 3 Results & discussion

Data Filtration and Strain Distribution

A total of 3,920 Helicobacter pylori sequences were downloaded from the BV-BRC and NCBI databases, of which 1,440 were complete sequences. The numbers of strains remaining after each filtration step are shown in Figure 1. The filtration step that removed the most strains was the CheckM completeness filter, as the completeness value was not provided for many sequences available on BV-BRC.

The remaining complete sequences are composed of strains from highly diverse geographic locations, including 64 countries from all major continents. The top three countries, the USA, Colombia, and China, account for approximately 25% of all samples, while sequences from the top 10 countries represent more than 50% of the total. This broad distribution ensures that the dataset reflects the global diversity of H. pylori strains, providing a robust basis for the pangenome analysis.

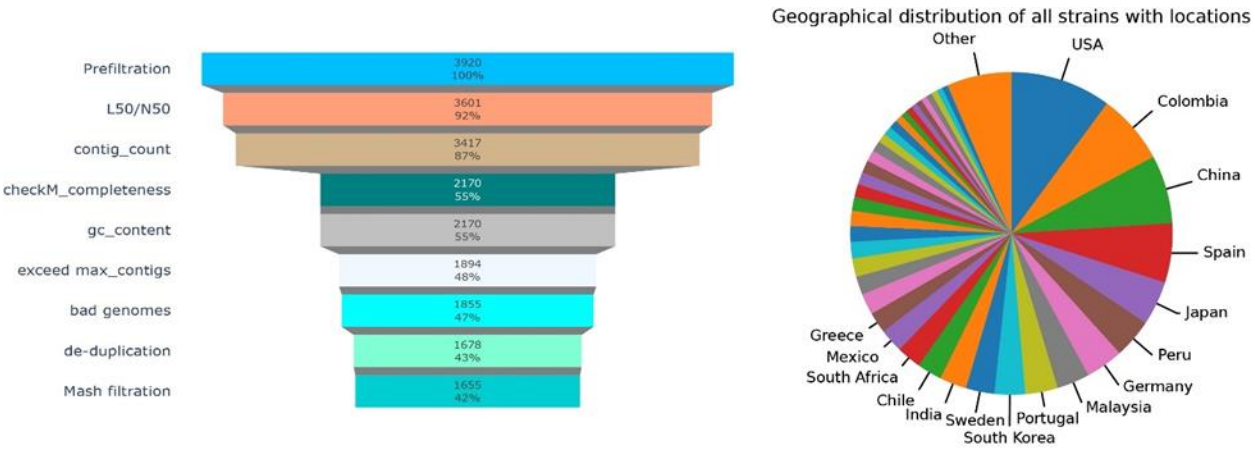

Figure 1. Numbers of strains after each filtration step and geographic distribution of strains

Pangenome and Gene Enrichment

Of the 1,655 strains that passed all filtration criteria, 1,341 complete sequences were selected for further CD-HIT clustering. The CD-HIT algorithm was run with an 80% identity threshold (see Supplementary Figure S1), and genes were classified into core, accessory, and rare genomes based on their frequency across the strains (Figure 2). Genes that appeared in fewer than 98 strains (7.3%) were categorized as rare genes, while those present in more than 1,304 strains (97.2%) were classified as core genes. The remaining genes were designated as accessory genes.

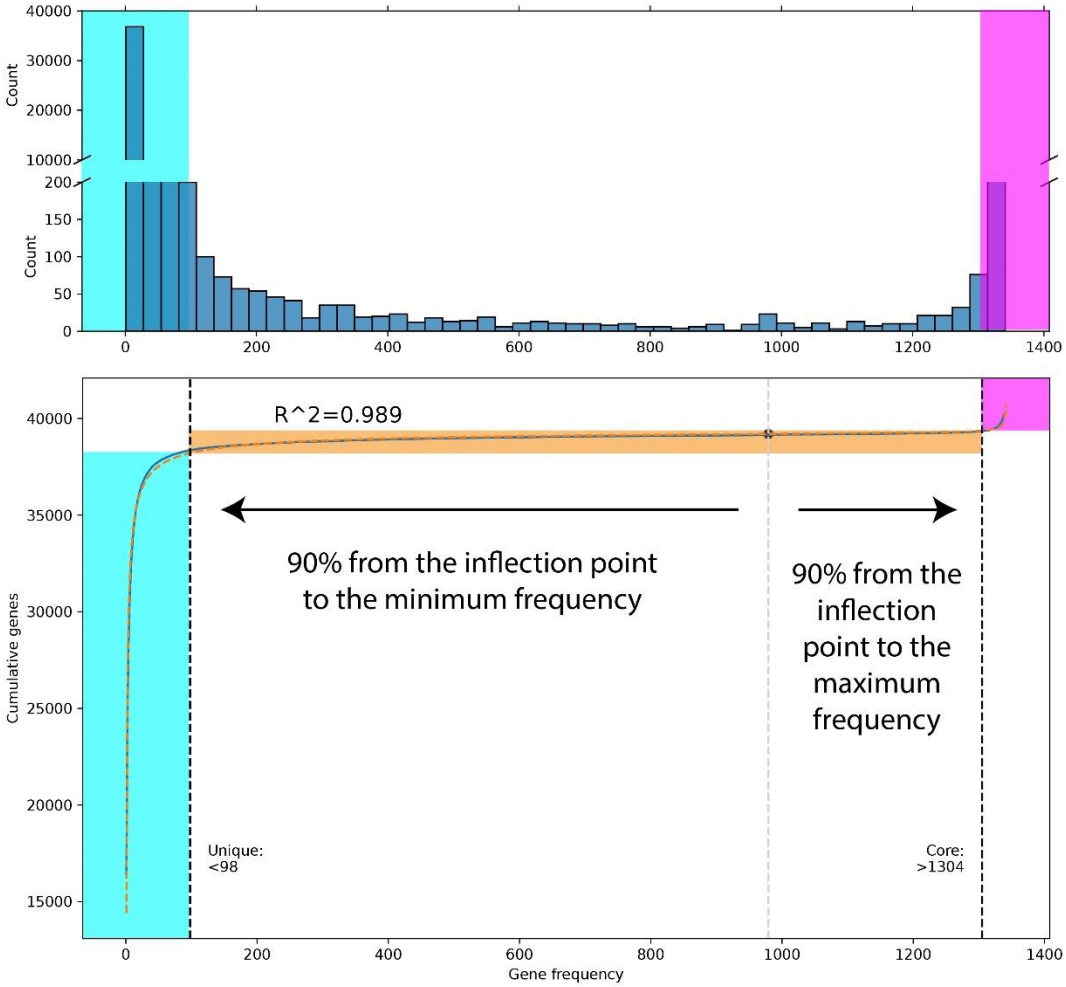

Figure 2. Gene frequency distribution and simulation curve fitted to the cumulative gene frequency with cutoff values for rare and core genes.

The results identified 1,015 core genes, 986 accessory genes, and 38,357 rare genes. Remarkably, despite using more than ten times the number of sequences compared to other studies, the number of core and accessory genes identified aligns closely with earlier findings. This includes a study that extrapolated the number of core genes to be approximately 1,111 using 56 H. pylori strains[36]. By fitting the gene frequency distribution to Heaps' law, it was confirmed that H. pylori has an open pangenome with a lambda value of 0.436. This suggests that the gene pool is not finite and will likely expand as more strains are added to the analysis.

For the core genome, it was not surprising to find that genes associated with essential functions such as translation, cell membrane biogenesis, and energy production comprised a significant portion of the core genome (Figure 3). In contrast, nearly half of the genes in the accessory genome were either unidentified by eggNOG or had unknown functions. Among those with known functions, the top categories included replication and repair, cell motility, and cell membrane biogenesis.

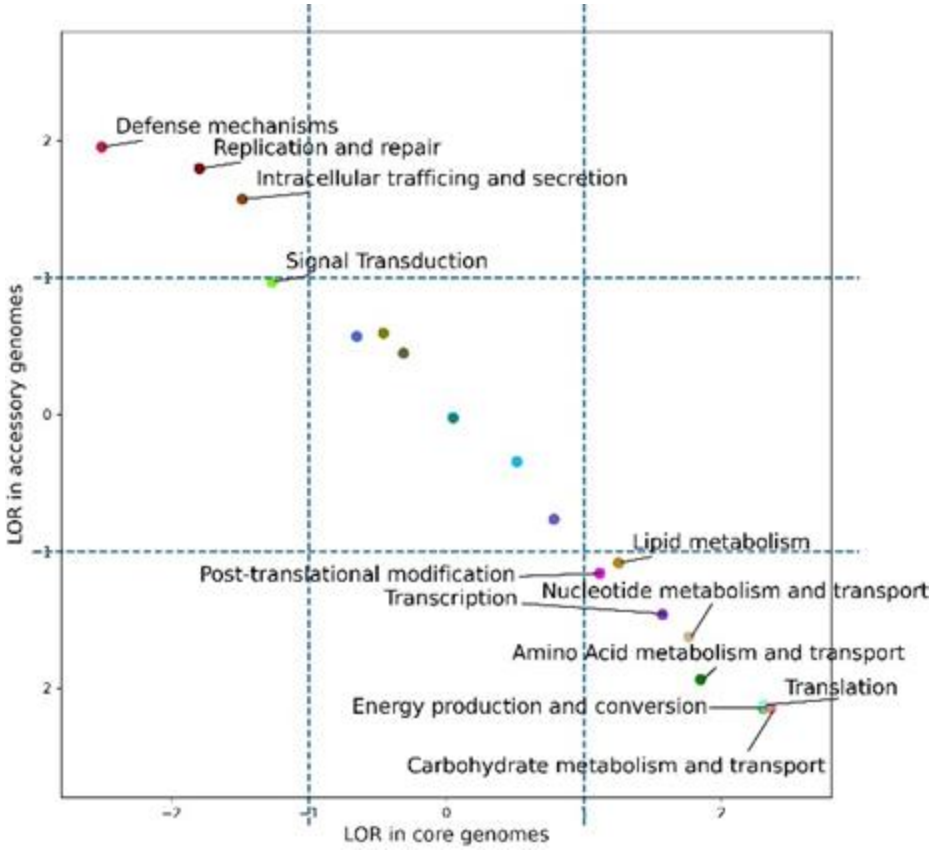

Figure 3. Log odds ratio comparison of COG categories between core and accessory genomes.

The COG enrichment analysis, performed using Fisher's exact test and log odds ratios (LOR) comparison, revealed that translation, carbohydrate metabolism, and energy production were the top three enriched categories in the core genomes. Meanwhile, intracellular trafficking and secretion, replication and repair, and defense mechanisms were the most enriched categories within the accessory genome. Notably, the LOR for the core and accessory genomes demonstrated an inverse linear relationship.

Mash clustering and NMF analysis

Mash clustering identified 59 robust clusters among the 1,341 H. pylori strains, with most of the larger clusters (those containing more than ten strains) originating from a single experiment or the same location (Figure 4 top). This clustering pattern confirms the distinct

geographic distribution of H. pylori strains. This outcome contrasts significantly with our earlier test on a sample of 428 strains, which resulted in only 15 clusters. We did not anticipate that quadrupling the sample size would also quadruple the number of Mash clusters. This discrepancy may be attributed to the Mash algorithm or the high variability present in H. pylori sequences.

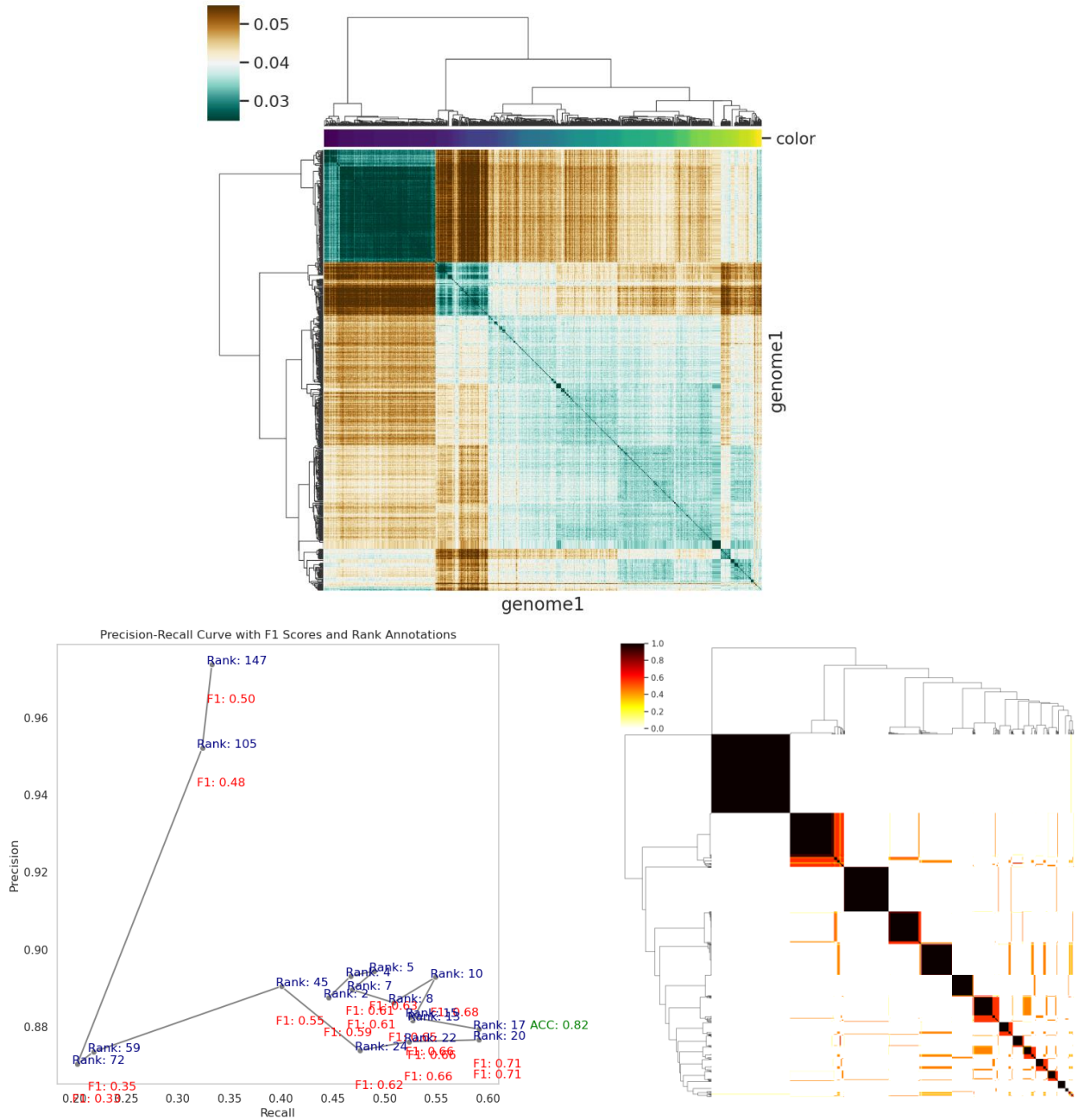

Figure 4. Mash clusterings (top), NMF performance at various rank (left), and NMF clustering (right).

Given these differences, it was necessary to reevaluate the optimal rank choice for non-negative matrix factorization (NMF) analysis. To address this, NMF analysis was performed across multiple ranks to determine the best configuration. Ultimately, a rank of 17 was selected based on the highest F1 score, which provided the most accurate reconstruction of the accessory gene matrix (Figure 4 bottom left).

Characterization of Phylogroups

We began by characterizing the 17 phylogroups using the A matrix with available metadata to gain insights into the phylon-specific traits and geographical distribution of these groups. Despite our efforts, we found no significant commonalities when categorizing the phylogroups based on Multi-Locus Sequence Typing (MLST) or virulence typing. This lack of commonality may stem from two main factors: the majority of sequences lack detailed metadata, and the high genetic diversity within H. pylori makes it difficult to draw clear connections. We turned to geographic information as a more reliable basis for characterizing and naming the phylogroups. The analysis revealed distinct patterns, with many strains from a single phylogroup originating from geographically close areas. This geographic clustering also reflects historical human migration patterns.

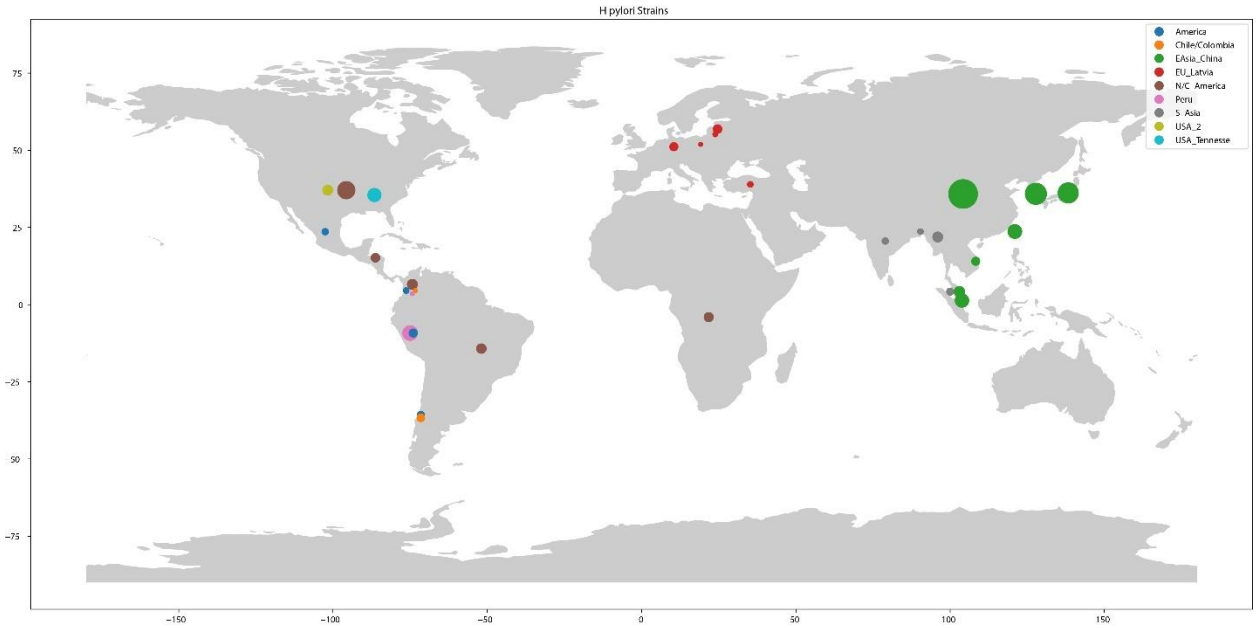

Figure 5. Geographic locations of strains in phylon identified.

After exploring the phylogroups, the focus shifted to characterizing the L matrix, aiming to identify genes that distinguish the phylogroups. It was observed that genes associated with membrane biosynthesis exhibit high variances across the phylogroups, suggesting they may play a crucial role in differentiating strains. This finding is particularly intriguing because membrane biosynthesis genes are among the top three enriched categories in the core genome but not in the accessory genome. The presence of such genes in the core genome underscores their essential role in bacterial survival and adaptability. Thus, the variability of these genes across phylogroups suggests that they might influence membrane protein secretions or interactions. Despite this potential significance, the exact functions of these genes remain unknown and require further research.

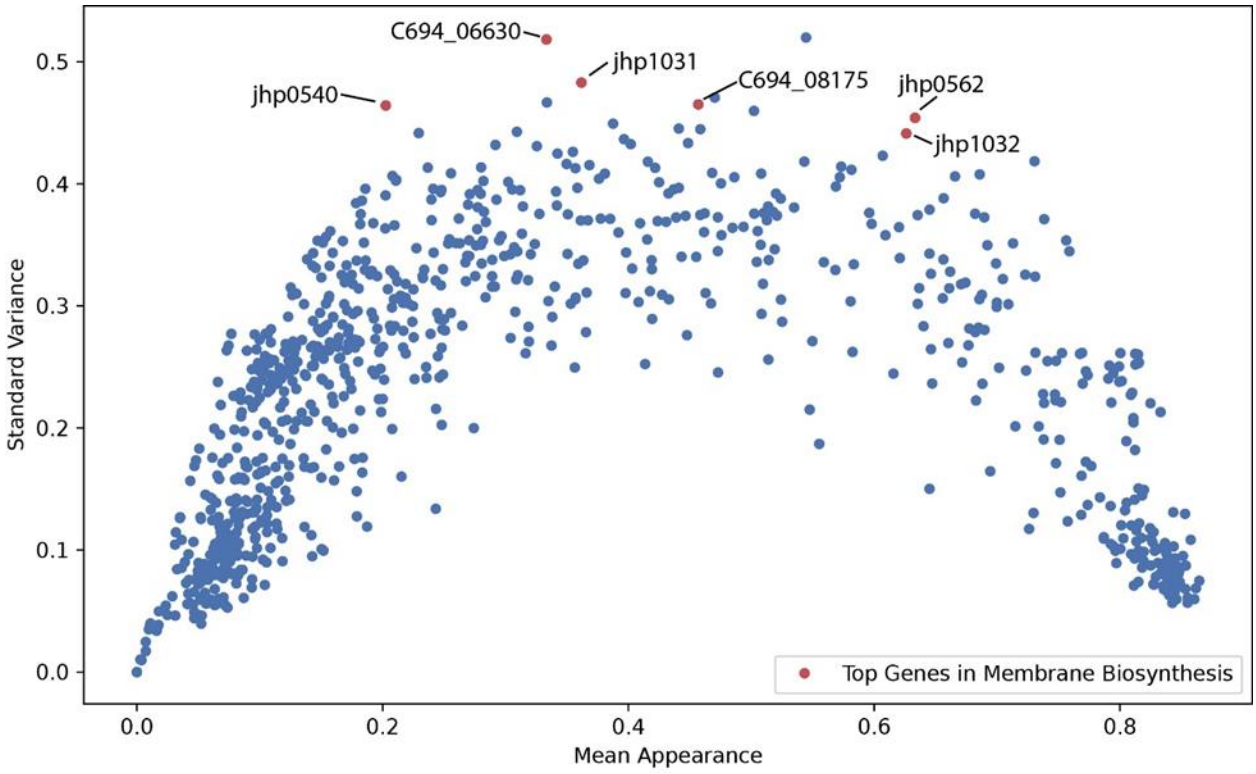

Figure 6. Gene variances vs mean appearances with top gene category annotated

Two distinct methods were employed to characterize the L matrix: a top-down approach and a random forest classifier. In the top-down approach, the binarized L matrix was hierarchically clustered to analyze phylogenetic relationships and identify exclusive genes belonging to each phylon. However, drawing conclusions about phylon-specific traits proved challenging because more than half of the exclusive genes remain unannotated. The analysis of highly variable genes hinted at the significance of genes related to the membrane. BLAST searches were employed to identify commonly expressed outer membrane proteins (OMPs) and their orthologs. By examining the presence of OMPs in the binarized L matrix, an OMP profiling table was created. In this profiling, v1 and v2 represent two orthologs, while intermediate colors indicate that both exist. This profiling approach can help identify similarities between closely related phylogroups. However, the top-down approach did not appear to be the most effective method for uncovering the genetic basis of phylogroups in the case of H. pylori.

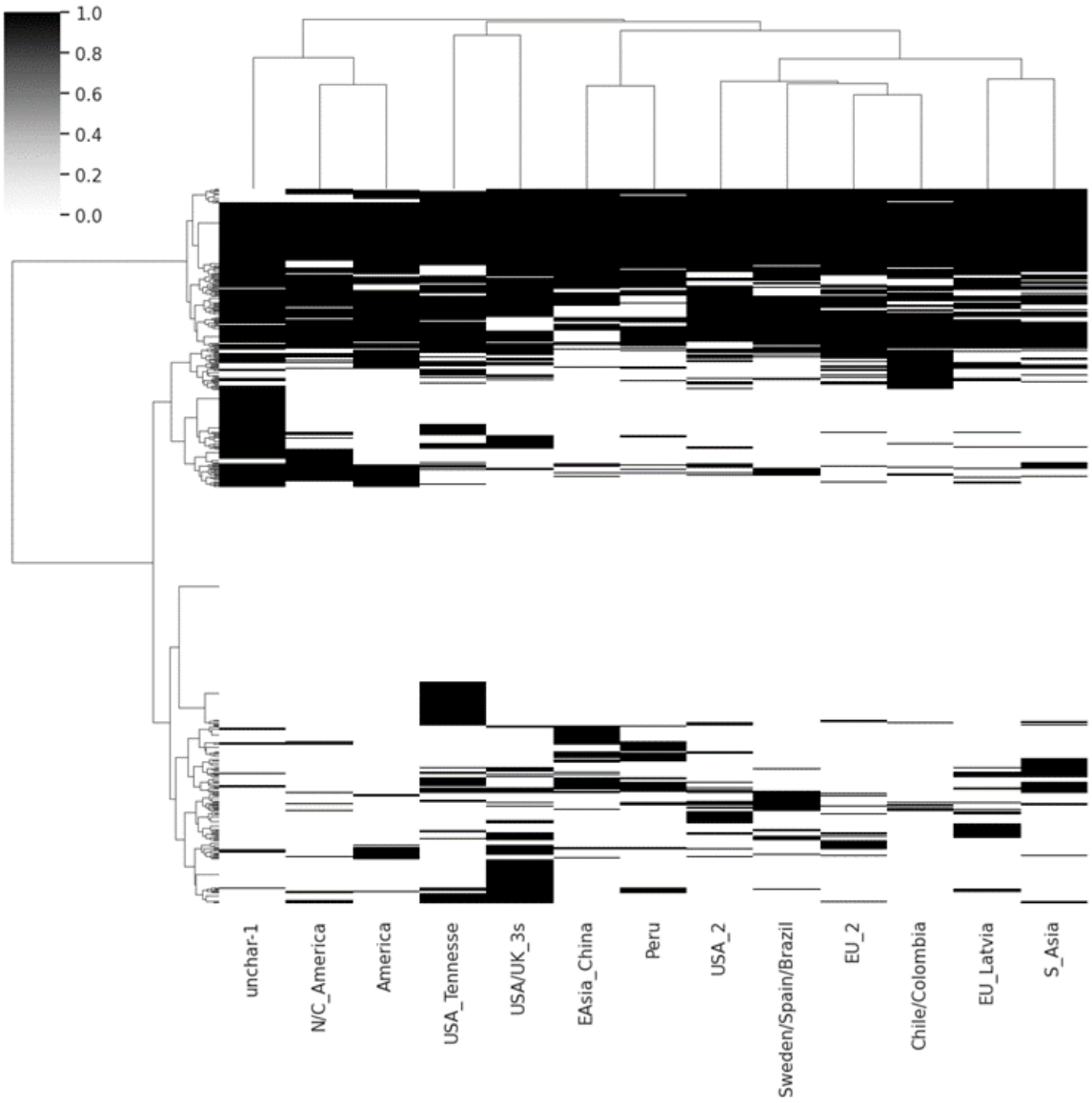

Figure 7. Ordered L matrix with removed uncharacterized phylon.

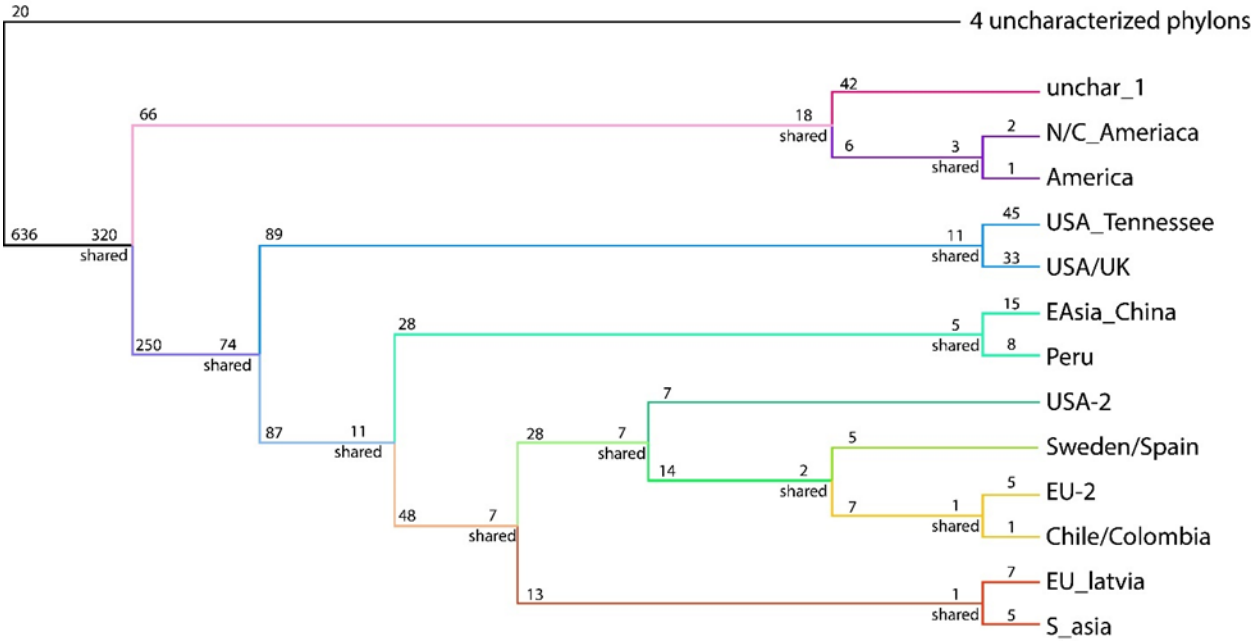

Figure 8. Numbers of exclusive genes after each phylon split.

To address the limitations of the top-down approach, a random forest classifier was employed to identify differential genes more automatically. This method allowed pinpointing genes with the highest feature importance when the classifier was trained to differentiate one subgroup from the rest. For example, when the random forest classifier was used to distinguish the USA_Tennessee and USA/UK phylogroups from the rest, one of the top genes identified was the fur gene. These two phylogroups possessed a unique ortholog of the fur gene compared to other phylogroups.

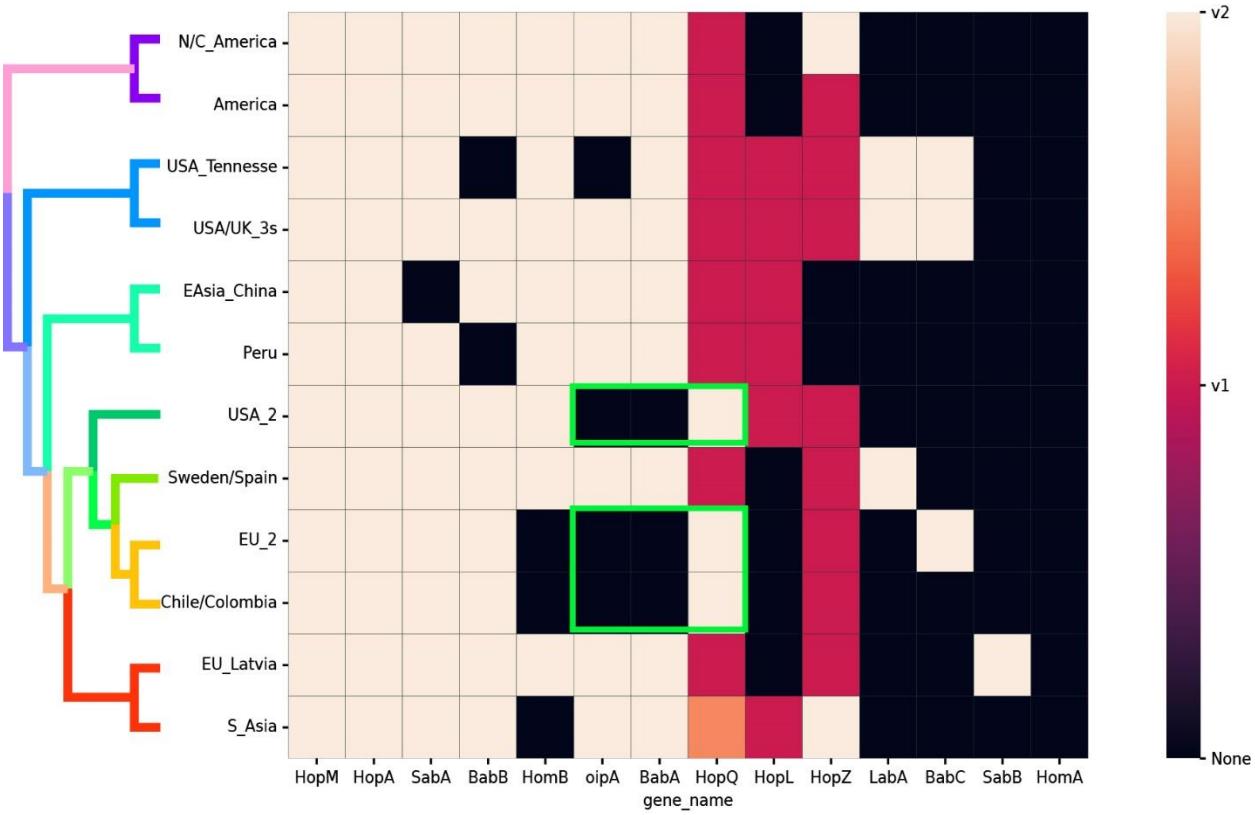

Figure 9. OMP profiles of all phylogroups.

The advantage of the random forest method lies in its ability to compare multiple phylogroups that are not under a single phylogenetic node to the rest. From the OMP profile, it was observed that four closely related phylogroups—USA_2, Sweden/Spain, EU_2, and Chile/Colombia—exhibit slightly different profiles in genes such as BabA, HopQ, and oipA. When the random forest classifier was run to compare USA_2, EU_2, and Chile/Colombia against the rest, these three genes emerged among the top 20 genes with high feature importance. The gene with the highest feature importance was vacA, a known virulence factor. Further analysis revealed that these three phylogroups possess the s2-subtype vacA, an avirulent version, while the Sweden/Spain phylogroup carries the virulent version. This observation aligns with the understanding that avirulent and virulent strains may have different OMPs, as OMP interaction with the host is a critical step in colonization and infection.

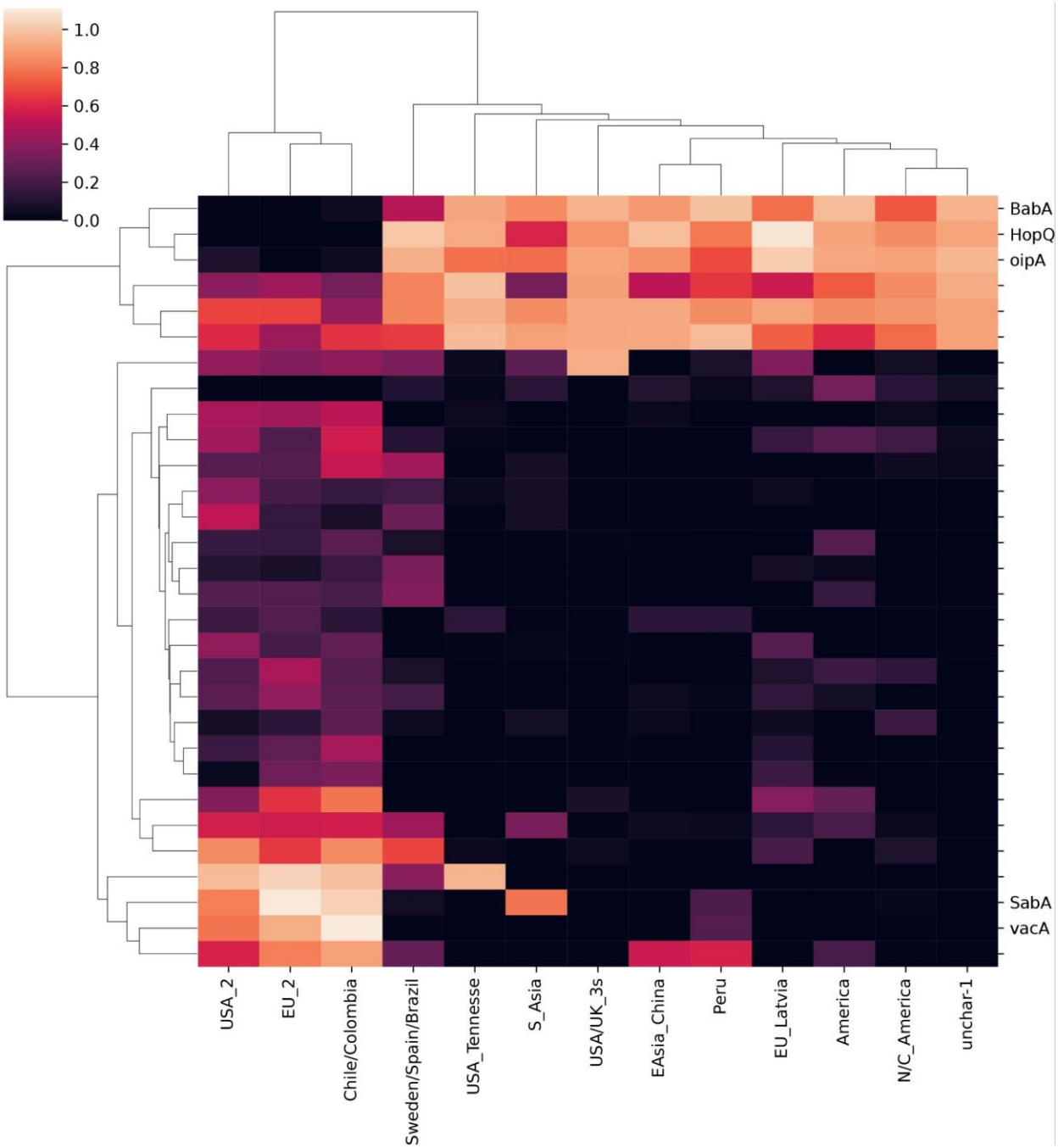

Figure 10. Average feature importance of genes identified by Random Forest Classifier when comparing USA_2, EU_2, and Chile/Colombia to the rest.

Overall, the random forest method proves useful, especially for poorly annotated genomes, by narrowing down target genes that contribute most significantly to phylon classification. Unlike the top-down approach, which relies on binarization and hierarchical

clustering, the random forest method allows for a more flexible and automated exploration of genetic diversity.

The use of machine learning was intended not for strain classification, but to identify components with the highest feature importance. To validate this approach, the reproducibility of feature identification using the random forest method was compared against other machine learning methods such as the linear SVM and simple decision tree. The random forest and the linear SVM methods demonstrated relative low deviation in feature importance across runs, but considering the linear SVM model's assumption of linear relationship and sensitivity to outliers, we chose the Random Forest Classifier with more flexibility and interpretability.

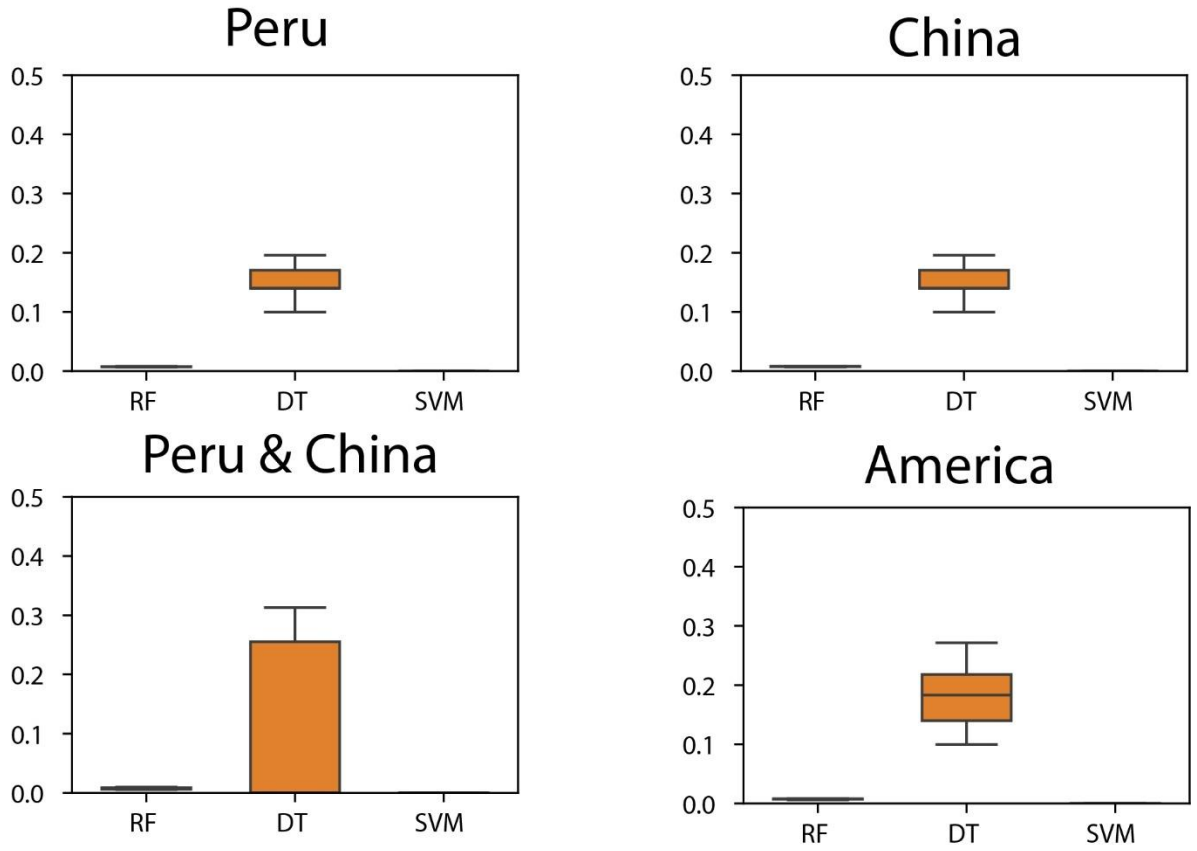

Figure 11. Standard deviation comparisons among three machine learning methods.

There are four phylogroups that remain uncharacterized. These groups contain a small set of genes with only 20 unique genes in total, which are suspected to be mobile elements of H.

pylori. Further research is required to understand the exact functions of these genes and confirm their role in H. pylori.

iModulon Analysis

Extensive RNA-seq data is often required to understand gene functionality and regulatory networks, which can be challenging for many less-studied bacterial species. To address this limitation, gene categories identified from pangenome analysis can help construct a multi-strain RNA-seq dataset, enhancing statistical power and comparability in RNA-seq analyses such as iModulon analysis.

For this study, RNA-seq data of core and accessory genes from three different H. pylori strains: 26695, G27, and P12 were combined. iModulon analysis was performed on RNA-seq data for every single strain, as well as on combined core genomes (C) and combined core plus accessory genomes (CA). The results showed that using only core genomes yielded an explained variance of 0.479, which is below the standard, whereas using the CA genome achieved a significantly higher explained variance of 0.919. Including all rare genes raises the total number of genes to around 3,000, making the expression matrix too sparse, thus affecting further analysis.

Table 1. Statistic summary of iModulon results on 5 datasets.

Strain	# of samples available	# of samples used	Number of iModulons	Explained Variance	Number of genes	Number of genes included
26695	122	85	34	0.846	1553	666(42,8%)
G27	30	26	20	0.997	1546	505(32.6%)
P12	40	26	18	0.97	1572	800(50.9%)
Core	-	137	29	0.479	988	293(29.3%)
C+A	-	137	47	0.919	1716	1043(60.8%)

By increasing the sample size and excluding rare genes, the noise was effectively reduced, improving the clarity of the gene expression data. For example, the genes in the ribosome iModulon (rpl) identified in the CA dataset were 84.37% in the translation category, compared to just 50% in the G27 single-strain dataset.

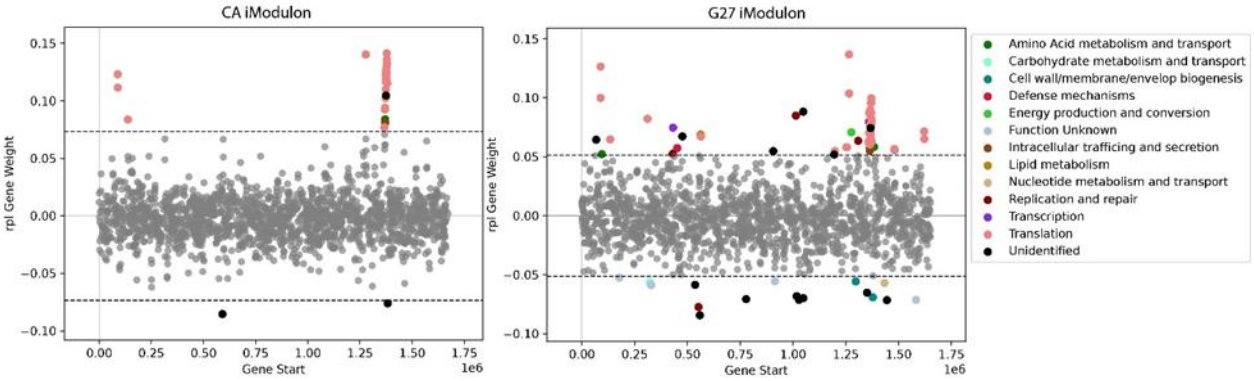

Figure 12. Identified ribosome iModulon (rpl) comparison between single-strain G27 and combined core plus accessory dataset.

Moreover, comparing iModulon activities across different strains enables drawing connections in regulatory networks, even when experiments are conducted on separate strains. In the pH_ure iModulon activity analysis, it was observed that this iModulon is upregulated under acidic conditions and downregulated when nickel is present with the wild-type NikR, suggesting that NikR plays a crucial role in pH response. With additional samples from different conditions and strains combined, a better understanding of the comprehensive global regulatory role of genes, such as NikR in H. pylori, can be achieved.

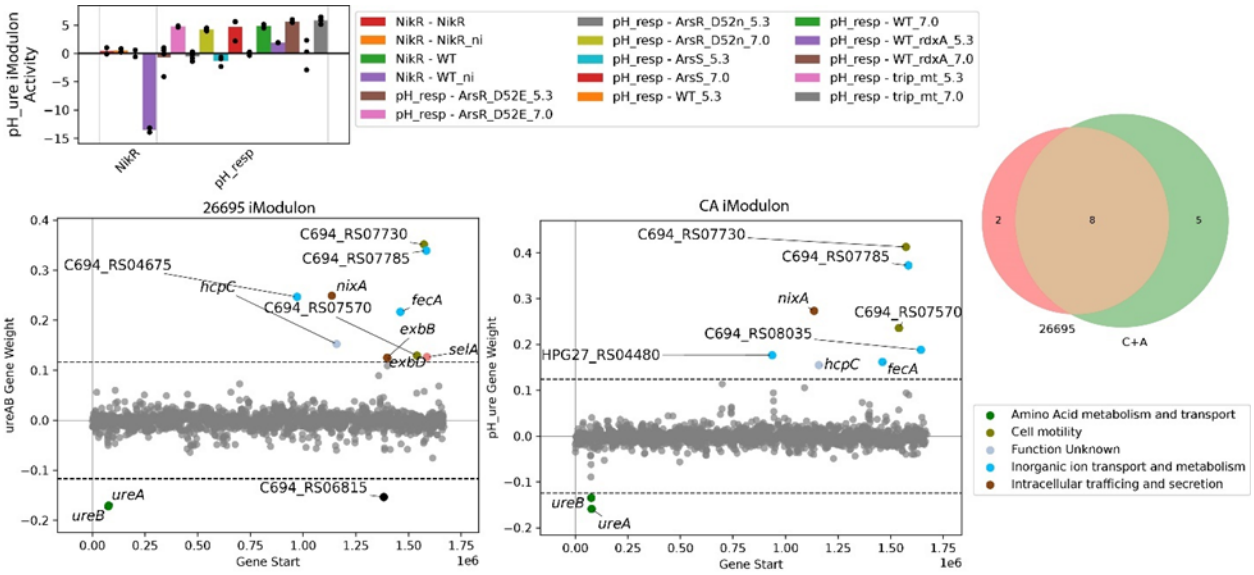

Figure 13. NikR-related iModulon activity comparison and Venn diagram for genes identifies in the two iModulons.

In the rpl iModulon activity comparison, the rpl activity was upregulated in tetracycline-treated samples and downregulated in the NikR knockout group, suggesting that NikR may also regulate ribosome synthesis. In E. coli, NikR interferes with translation through toxin-antitoxin systems involving the mccB and mccC genes. This may also be the case in H. pylori, as it possesses orthologs of these genes.

The construction of a multi-strain RNA-seq dataset using pangenome analysis enhances the reliability and interpretability of iModulon analysis. This approach provides a more robust understanding of the regulatory networks and gene functionality in H. pylori, paving the way for further research into its pathogenic mechanisms.

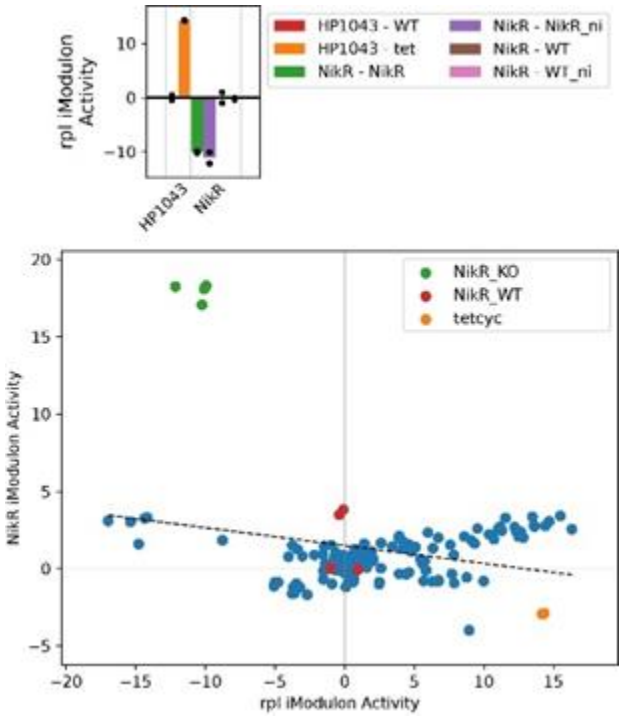

Figure 14. iModulon activity comparison between NikR and ribosome iModulon.

Limitation

A few limitations must be acknowledged as these offer opportunities for future research and refinement of methodologies to better understand this complex pathogen. A major limitation is the lack of functional annotations for many of the genes identified, especially within the accessory and rare genomes. This gap in knowledge hinders our ability to fully comprehend the roles of these genes in pathogenicity and adaptability.

Additionally, this study relies heavily on computational algorithms such as CD-HIT, Non-negative matrix factorization (NMF), and random forest Classifiers. While these tools are powerful for data analysis, their inherent limitations can influence the outcomes. For example, the choice of parameters in NMF and the reliance on pre-defined thresholds in CD-HIT could affect the accuracy of gene clustering and phylogenetic grouping. These parameters might lead to

biases in identifying gene functions and strain relationships. Further refinement and validation of these methods could enhance their reliability and applicability, ensuring that the results accurately reflect the genetic architecture of the pangenome.

REFERENCES

1. Whalen MB, Massidda O. Helicobacter pylori: enemy, commensal or, sometimes, friend? *The Journal of Infection in Developing Countries*. 2015;9(06):674-678. doi:10.3855/jidc.7186

2. Pellicano R, Ribaldone DG, Fagoonee S, Astegiano M, Saracco GM, Mégraud F. A 2016 panorama of Helicobacter pylori infection: key messages for clinicians. *Panminerva Med*. 2016;58(4):304-317.

3. Hunt R h., Sumanac K, Huang JQ. Review article: should we kill or should we save Helicobacter pylori? *Alimentary Pharmacology & Therapeutics*. 2001;15(s1):51-59. doi:10.1046/j.1365-2036.2001.00107.x

4. Aljaberi HSM, Ansari NK, Xiong M, Peng H, He B, Wang S. Current Understanding of the Transmission, Diagnosis, and Treatment of H. pylori Infection: A Comprehensive Review. *IJMPD*. 2023;7(2):01-26. doi:10.22161/ijmpd.7.2.1

5. Malfertheiner P, Selgrad M, Bornschein J. Helicobacter pylori: clinical management. *Curr Opin Gastroenterol*. 2012;28(6):608-614. doi:10.1097/MOG.0b013e32835918a7

6. Kotilea K, Bontems P, Touati E. Epidemiology, Diagnosis and Risk Factors of Helicobacter pylori Infection. *Adv Exp Med Biol*. 2019;1149:17-33. doi:10.1007/5584_2019_357

7. Daryani NE, Taher M, Shirzad S. Helicobacter Pylori infection: A review. *Archives of Clinical Infectious Diseases*. Published online 2011. Accessed August 7, 2024. https://www.semanticscholar.org/paper/Helicobacter-Pylori-infection%3A-A-review-Daryani-Taher/c10c966ea5d2f02eed9a713db1bbe6cddf3a548d

8. McGowan CC, Cover TL, Blaser MJ. Helicobacter pylori and gastric acid: biological and therapeutic implications. *Gastroenterology*. 1996;110(3):926-938. doi:10.1053/gast.1996.v110.pm8608904

9. Krzyżek P, Grande R, Migdał P, Paluch E, Gościniak G. Biofilm Formation as a Complex Result of Virulence and Adaptive Responses of Helicobacter pylori. *Pathogens*. 2020;9(12):1062. doi:10.3390/pathogens9121062

10. Bani-Hani KE. The current status of Helicobacter pylori. *Saudi Med J*. 2002;23(4):379-383.

11. Wen Y, Marcus EA, Matrubutham U, Gleeson MA, Scott DR, Sachs G. Acid-Adaptive Genes of Helicobacter pylori. *Infect Immun*. 2003;71(10):5921-5939. doi:10.1128/IAI.71.10.5921-5939.2003

12. Karkhah A, Ebrahimpour S, Rostamtabar M, Koppolu V, Darvish S, Vasigala VKR, Validi M, Nouri HR. Helicobacter pylori evasion strategies of the host innate and adaptive immune responses to survive and develop gastrointestinal diseases. *Microbiol Res*. 2019;218:49-57. doi:10.1016/j.micres.2018.09.011

13. Díaz P, Valenzuela Valderrama M, Bravo J, Quest AFG. Helicobacter pylori and Gastric Cancer: Adaptive Cellular Mechanisms Involved in Disease Progression. *Front Microbiol*. 2018;9:5. doi:10.3389/fmicb.2018.00005

14. Eckloff BW, Podzorski RP, Kline BC, Cockerill FR. A comparison of 16S ribosomal DNA sequences from five isolates of Helicobacter pylori. *Int J Syst Bacteriol*. 1994;44(2):320-323. doi:10.1099/00207713-44-2-320

15. Golicz AA, Batley J, Edwards D. Towards plant pangenomics. *Plant Biotechnol J*. 2016;14(4):1099-1105. doi:10.1111/pbi.12499

16. Aggarwal SK, Singh A, Choudhary M, Kumar A, Rakshit S, Kumar P, Bohra A, Varshney RK. Pangenomics in Microbial and Crop Research: Progress, Applications, and Perspectives. *Genes*. 2022;13(4):598. doi:10.3390/genes13040598

17. Anani H, Zgheib R, Hasni I, Raoult D, Fournier PE. Interest of bacterial pangenome analyses in clinical microbiology. *Microb Pathog*. 2020;149:104275. doi:10.1016/j.micpath.2020.104275

18. Rasko DA, Rosovitz MJ, Myers GSA, Mongodin EF, Fricke WF, Gajer P, Crabtree J, Sebaihia M, Thomson NR, Chaudhuri R, Henderson IR, Sperandio V, Ravel J. The Pangenome Structure of *Escherichia coli* : Comparative Genomic Analysis of *E. coli* Commensal and Pathogenic Isolates. *J Bacteriol*. 2008;190(20):6881-6893. doi:10.1128/JB.00619-08

19. Golicz AA, Bayer PE, Bhalla PL, Batley J, Edwards D. Pangenomics Comes of Age: From Bacteria to Plant and Animal Applications. *Trends Genet*. 2020;36(2):132-145. doi:10.1016/j.tig.2019.11.006

20. Ali A, Naz A, Soares SC, Bakhtiar M, Tiwari S, Hassan SS, Hanan F, Ramos R, Pereira U, Barh D, Figueiredo HCP, Ussery DW, Miyoshi A, Silva A, Azevedo V. Pan-Genome Analysis of Human Gastric Pathogen *H. pylori* : Comparative Genomics and Pathogenomics Approaches to Identify Regions Associated with Pathogenicity and Prediction of Potential Core Therapeutic Targets. *BioMed Research International*. 2015;2015:1-17. doi:10.1155/2015/139580

21. Van Vliet AHM. Use of pan-genome analysis for the identification of lineage-specific genes of *Helicobacter pylori*. Boden R, ed. *FEMS Microbiology Letters*. 2017;364(2):fnw296. doi:10.1093/femsle/fnw296

22. Cao DM, Lu QF, Li SB, Wang JP, Chen YL, Huang YQ, Bi HK. Comparative Genomics of *H. pylori* and Non-Pylori *Helicobacter* Species to Identify New Regions Associated with Its Pathogenicity and Adaptability. *BioMed Research International*. 2016;2016:1-15. doi:10.1155/2016/6106029

23. Uchiyama I, Albritton J, Fukuyo M, Kojima KK, Yahara K, Kobayashi I. A Novel Approach to Helicobacter pylori Pan-Genome Analysis for Identification of Genomic

Islands. Cloeckaert A, ed. *PLoS ONE*. 2016;11(8):e0159419. doi:10.1371/journal.pone.0159419

24. Olson RD, Assaf R, Brettin T, Conrad N, Cucinell C, Davis JJ, Dempsey DM, Dickerman A, Dietrich EM, Kenyon RW, Kuscuoglu M, Lefkowitz EJ, Lu J, Machi D, Macken C, Mao C, Niewiadomska A, Nguyen M, Olsen GJ, Overbeek JC, Parrello B, Parrello V, Porter JS, Pusch GD, Shukla M, Singh I, Stewart L, Tan G, Thomas C, VanOeffelen M, Vonstein V, Wallace ZS, Warren AS, Wattam AR, Xia F, Yoo H, Zhang Y, Zmasek CM, Scheuermann RH, Stevens RL. Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. *Nucleic Acids Res*. 2023;51(D1):D678-D689. doi:10.1093/nar/gkac1003

25. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Connor R, Funk K, Kelly C, Kim S, Madej T, Marchler-Bauer A, Lanczycki C, Lathrop S, Lu Z, Thibaud-Nissen F, Murphy T, Phan L, Skripchenko Y, Tse T, Wang J, Williams R, Trawick BW, Pruitt KD, Sherry ST. Database resources of the national center for biotechnology information. *Nucleic Acids Res*. 2022;50(D1):D20-D26. doi:10.1093/nar/gkab1112

26. Thorell K, Muñoz-Ramírez ZY, Wang D, Sandoval-Motta S, Boscolo Agostini R, Ghirotto S, Torres RC, Falush D, Camargo MC, Rabkin CS. The Helicobacter pylori Genome Project: insights into H. pylori population structure from analysis of a worldwide collection of complete genomes. *Nat Commun*. 2023;14(1):8184. doi:10.1038/s41467-023-43562-y

27. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*. 2016;17(1):132. doi:10.1186/s13059-016-0997-x

28. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 2020;17(3):261-272. doi:10.1038/s41592-019-0686-2

29. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 2012;28(23):3150-3152. doi:10.1093/bioinformatics/bts565

30. Hyun JC, Monk JM, Palsson BO. Comparative pangenomics: analysis of 12 microbial pathogen pangenomes reveals conserved global structures of genetic and functional diversity. *BMC Genomics*. 2022;23(1):7. doi:10.1186/s12864-021-08223-8

31. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30(14):2068-2069. doi:10.1093/bioinformatics/btu153

32. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the

Metagenomic Scale. *Molecular Biology and Evolution*. 2021;38(12):5825-5829. doi:10.1093/molbev/msab293

33. Chauhan SM, Ardalani O, Hyun JC, Monk JM, Phaneuf PV, Palsson BO. The Pangenome of Escherichia coli. Published online June 8, 2024:2024.06.07.598014. doi:10.1101/2024.06.07.598014

34. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011;12(85):2825-2830.

35. Rychel K, Decker K, Sastry AV, Phaneuf PV, Poudel S, Palsson BO. iModulonDB: a knowledgebase of microbial transcriptional regulation derived from machine learning. *Nucleic Acids Res*. 2021;49(D1):D112-D120. doi:10.1093/nar/gkaa810

36. Gressmann H, Linz B, Ghai R, Pleissner KP, Schlapbach R, Yamaoka Y, Kraft C, Suerbaum S, Meyer TF, Achtman M. Gain and Loss of Multiple Genes During the Evolution of Helicobacter pylori. *PLoS Genet*. 2005;1(4):e43. doi:10.1371/journal.pgen.0010043