Analyzing and Interpreting Bacterial Omics Data for Antibiotic and Vaccine Development

Mohsin Ali, Jitendra Choudhary

***Abstract*—Motifs are short, recurring nucleotide sequences in DNA that play a crucial role in gene regulation, transcription, and genomic stability. This study adds to our knowledge of bacterial genomics by looking at differences in nucleotide sequences found in bacterial omics data. It focuses on important patterns like the TATA Box, CAAT Box, Start Codon, Microsatellite Motifs, EcoRI recognition site, and CpG Islands. Some of the most powerful bioinformatics tools we use are k-mer analysis, truncated SVD, and t-SNE to look at the structural differences in these important genomic parts. This study shows how these differences affect the functionality and adaptability of bacteria. It provides us with information about how microbes have changed over time and suggests possible targets for antimicrobial strategies. Statistical tests, like the Kruskal-Wallis H-test and the pairwise Mann-Whitney U-test, are used to figure out how important it is that different groups of bacteria have different motif frequencies. This discovery gives us a new way to look at how complicated the regulation of bacterial genomes is.**

***Index Terms*—omic data, Vaccine, Nucleotide Sequence, Statistical tool, t-SNE.**

I. Introduction

THE study of bacterial omics data involves examining the genetic material of microorganisms to understand their genetic design and growth. This field has revolutionized our understanding of microbial organisms and has provided new insights into the mechanisms underlying microbial infections, antibiotic resistance, and pathogenesis. Additionally, these omic technologies integrate with statistical methods to develop new tools for understanding complex biological data [1-2].

Nucleotide sequence variation means variation in bacteria's genetic structure. Furthermore, these genetic structure differences occur within and between species from genetic processes like modification, duplication, and horizontal gene transformation[3-4]. Moreover, to comprehend the distinctions and their interdependence, we need to examine the impact of harmful bacteria and resistance[5-7].

The nucleotide sequences of bacteria, through omic technology, give you insight and meaningful knowledge of characteristics and adaptation. Circular consensus sequencing (CCS) makes long-read sequencing systems more accurate, which lets scientists find single-nucleotide differences across whole genes. The 16S rRNA genes are one of the key essential marker genes for both classifying and identifying bacteria sequences [8-10]. Due to the rapid growth of bacterial genome sequencing, numerous tools are now available for genomic data analysis. However, each tool has different algorithms, user interfaces, hardware requirements, and programming languages. Integrating these tools is critical for interpreting any genomic sequencing data. Recent research has linked specific genes found in the gut flora to inflammation. To find out how different these bacterial genes are, researchers used both deep sequencing and linotyping, a data analysis tool that finds places where mutations happen a lot in short pieces of DNA. Amplicon sequencing was performed to amplify the pks island, tcpC, and usp genes, all of which are associated with inflammation and carried by specific strains of E. coli[11].

Working with massive data sets and complex computer systems can lead to problems associated with accessibility, repeatability, and transparency. The Orione framework [12-14], [23-24] represents a significant development in resolving the above problems. The framework also brings together open-source bioinformatics tools created explicitly for microbiologists. As a result, it will be easier to conduct data-intensive analyses and quality control. The bacterial genomic study is continuing to evolve and become more complex. Whon et al.'s survey from 2021 [15-17] highlights the value of multi-omics techniques in understanding the relationships between and functions of microbes. These integrated methods are necessary to uncover the complex connections within microbial ecosystems. Furthermore, this study has important implications for understanding bacterial genetics and ecology. All omic technologies [18-20] have expanded microbial community structure beyond culture-based methods. These days, microbial community research increasingly uses multi-omics techniques compared to single-omics analysis.

The complexity of multi-omics data requires advanced computational and analytical approaches [25-26]. These include network fusion, matrix factorization, and factor analysis [21]. The limitation of existing work is that no researcher has reviewed this bacterial omic data and has not interpreted it using k-mer analysis with t-SNE. This data provides insight into the complex relationship between sequence variation and the function of bacteria and affects the bacteria cluster.

The Kmer algorithm finds application in many areas of genomics, including metagenomics classification of genes, bacterial bioinformatics, and more [27-28]. In my research, we use this algorithm to analyze bacterial omic data and interpret slight differences in nucleotide sequence variability. The Kmer [22] algorithm converts these differences into numerical counts, which are then used for clustering and statistical modeling. We aim to examine the complex relationship between sequence variation and bacterial functionality. In addition, our theory says that some nucleotide sequences form motifs that show how biological functions are not evenly distributed among bacterial clusters. We also use the Kruskal-Wallis H-test and the pairwise Mann-Whitney U-test to look at the data and find patterns that repeat and significant

Manuscript received October 24, 2024; revised March 22, 2025.

Mohsin Ali is an Assistant Professor of Computer Science, Medicaps University, Indore 453331, INDIA (corresponding author to provide email: coolbuddy.next.door@gmail.com).

Jitendra Choudhary is an Associate Professor of Computer Science, Medicaps University, Indore 453331, INDIA (e-mail:jitendra.scsit@gmail.com).

differences between the groups. This data can help develop new antibiotics and vaccines for diseases caused by bacteria.

NuII Hypothesis (H_0) - Let $X_{i1}, X_{i2}, \ldots, X_{\mathrm{in}}$ represent the nucleotide sequences from the i^{th} bacterial cluster, where each X_{ij} is a vector of nucleotide frequencies or other derived features.

Define M_{ik} as the frequency of the k^{th} motif or sequence pattern in the i^{th} cluster.

The hypothesis states that there is no significant difference in the distribution of these sequences and motifs among clusters:

$$\mathrm{Var}\,(X_{1j}) = \mathrm{Var}\,(X_{2j}) = \cdots = \mathrm{Var}\,(X_{nj})$$
$$M_{1k}{=}M_{2k}{=}\cdots{=}M_{nk}$$

This implies that both the variability of sequences and the specific frequencies of motifs are constant across all clusters. Alternative Hypothesis (H_1)

The hypothesis suggests significant variability in nucleotide sequences among different clusters, particularly in the distribution of certain motifs or patterns.

Mathematically, this is expressed as at least one variance or motif frequency being different:$\exists i, j$ such that $\mathrm{Var}\,(X_{ij}) \neq \mathrm{Var}\,(X_{kj})$ for some k or $M_{ik} \neq M_{jk}$ for some j, k

This means that at least one cluster shows a distinct pattern in either the overall sequence variability.

Two key questions drive our research: What patterns of nucleotide sequence variability can be observed across different bacterial clusters? How do these variations correlate with specific biological functions or characteristics within bacteria? To find the answers, we use k-mer-based clustering and statistical modeling to look at the differences in nucleotide sequences in bacterial omics data. We also look for and study how key nucleotide motifs are spread across different bacterial clusters and try to figure out what these differences mean for function. Our goal is to contribute to a deeper understanding of bacterial genomics.

II. Method Details

A. Flow Diagram

We employ a flow diagram to illustrate the step-by-step process of calculating each motif sequence output. Data pre-processing, dimension reduction, statistical analysis, and visualization are utilized to analyze the motif sequence in an efficient manner as shown in Fig. 1.

B. Tools

We studied the nucleotide sequence variability in bacterial omics using advanced technology. Google Colab Pro, equipped with 52 GB of RAM, facilitated the research. A T4 GPU processor assisted in achieving the study outcomes. Using cutting-edge technology in this research ensured the findings were accurate and reliable.

C. Hypothesis Assumption

We evaluate the hypotheses using the Kruskal-Wallis H-test, a non-parametric method suitable for our analysis. The test statistic for the Kruskal-Wallis test is defined as follows: The p-value obtained from this test is less than our chosen significance level, α = 0.05. In that case, we reject the null hypothesis, indicating a significant variability in nucleotide sequences between clusters. The test statistic for the Kruskal-Wallis test is defined as follows:

$$H{=}\frac{12}{N(N{+}1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i}{-}3(N{+}1)$$

Where:

1) N is the total number of observations across all groups.
2) k is the number of groups (clusters).
3) R_i is the sum of ranks in the i^{th} group.
4) n_i is the number of observations in the i^{th} group.

D. Data collection

Our study used the NCBI dataset to analyze nucleotide sequences in bacteria. We used the 'bacteria.1.1.genomic.fna' file, which comprises genomic sequences from various bacterial species and is available on the NCBI website. Fig. 2 shows the conversion of the .fna file to CSV format, which is easy to analyze, and Fig. 3 shows the data information of the .fna file, including sequence and ID.

E. Sequence Processing

In the pre-processing section, the first pre-processing step is to convert the sequential character of omic data into numerical form, which machine learning algorithms can understand. We utilized the k-mers algorithm, available in the feature_extraction sklearn library, as shown in Fig. 4. This algorithm transforms each sequence into a concatenated string, which is then converted into a numerical representation using the CountVectorizer function, also shown in Fig. 4. Lastly, the resultant numerical sequence via k-mers is 26370 * 1156306.

After converting the nucleotide sequences numerically, the next step involves performing truncated SVD. We optimized the nucleotide sequence data to enhance machine learning efficiency using truncated SVD, which reduces dimension without losing important details. Also, truncated SVD in our model is a smart way to deal with the huge amount of nucleotide sequence data and makes the results more accurate and better, as seen in Fig. 5. After reducing the number of dimensions, we used t-distributed Stochastic Neighbor Embedding (t-SNE) to see the omic data. The t-SNE plot in Fig. 6 is a two-dimensional data representation. This visualization provides easily understood data patterns and clusters that may not be clear in the high-dimensional space. Additionally, the plot shows that the use of truncated SVD with t-SNE effectively reveals the complex biological data.

F. Clustering algorithm on t-SNE results

We used the KMeans clustering algorithm to precisely define these groups and study their biological significance. This split the data into ten separate clusters (k=10), as seen in Fig. 7. Furthermore, this approach enabled us to quantitatively analyze the data structure that t-SNE qualitatively unveiled. Every cluster corresponds to different underlying biological states or conditions, which are key in this study.

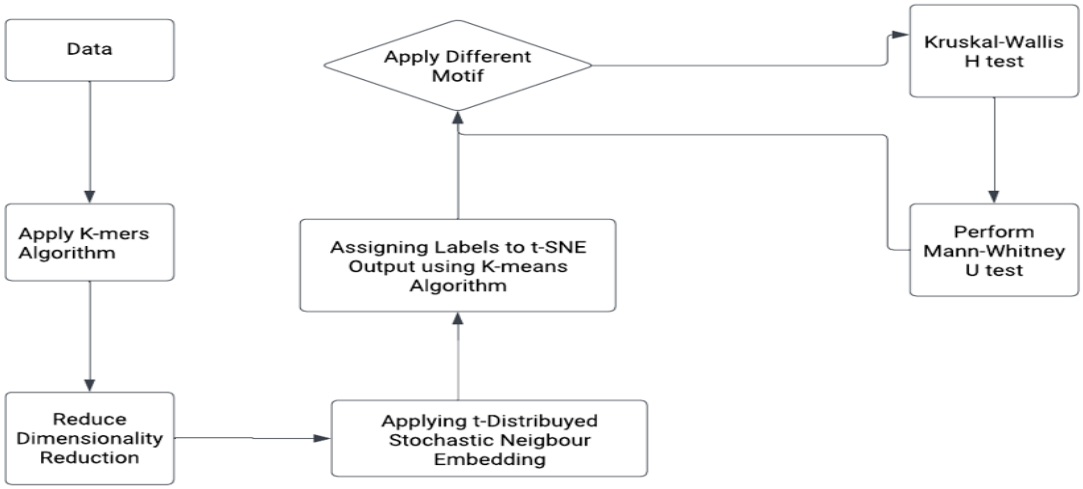

Fig. 1. Flow Diagram(Self Made)

```python
from Bio import SeqIO
import pandas as pd

# Define the path to your FastA file
fasta_file = 'G:/bacteria.1.1.genomic.fna'

# Create an empty list to hold the data
data = []

# Use Biopython to read the FastA file
for record in SeqIO.parse(fasta_file, "fasta"):
    identifier = record.id
    sequence = str(record.seq)
    data.append([identifier, sequence])

# Convert the list to a DataFrame
df = pd.DataFrame(data, columns=['ID', 'Sequence'])

# Save the DataFrame to CSV
csv_file = 'omic.csv'
df.to_csv(csv_file, index=False)
```

Fig. 2. Conversion of fna file to csv (Self Made)

```python
import pandas as pd

# Reload the dataset
path = "/content/drive/My Drive/omic.csv"
data = pd.read_csv(path)

# Displaying the first few rows of the dataset to understand its structure
data.head()
```

	ID	Sequence
0	NZ_QXLC01000017.1	AAGAAGACGAAAAGCAATGAGACGTAAAGTCTCACTGGTAATCGCA...
1	NZ_QXLC01000018.1	CCAAAAGAAATTTATAGAGCAAACTCGATAGATAAGGCGATGATGA...
2	NZ_QXLC01000019.1	AAGGTATAGTTAGTTACTGTATCACCTGCTTTAGGTAATATGGGTT...
3	NZ_QXLC01000020.1	GGGAAAGTAATCGGTTGGGTTGATACTCGTGCACTCGATACGTTCT...
4	NZ_QXTX01000001.1	CTAATCCTTCAGCAGTTTTAATCACTTTTTCAAGTTCTGATTTATC...

Fig. 3. Data Information

```python
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.manifold import TSNE
import matplotlib.pyplot as plt

# Function to convert sequences into k-mers
def get_kmers(sequence, k=10):
    return [sequence[x:x+k].lower() for x in range(len(sequence) - k + 1)]

# Applying the function to the dataset
data['kmers'] = data['Sequence'].apply(lambda x: ' '.join(get_kmers(x)))

# Using CountVectorizer to convert the sequences into numerical data
vectorizer = CountVectorizer()
X = vectorizer.fit_transform(data['kmers'])
```

```python
X.shape
```

```
(26370, 1156306)
```

Fig. 4. Applying k-mers Algorithm (Self Made)

```python
# Standardize the features

from sklearn.preprocessing import StandardScaler
scaler = StandardScaler(with_mean=False)
X_scaled = scaler.fit_transform(X)

# Apply SVD to reduce dimensionality
from sklearn.decomposition import TruncatedSVD

# Applying Truncated SVD
svd = TruncatedSVD(n_components=3)
X_svd = svd.fit_transform(X_scaled)
```

```python
X_svd.shape
```

```
(26370, 3)
```

Fig. 5. TSVD (Self Made)

G. Data Statistics

In this data, these statistics show each cluster's mean, standard deviation, and other properties in Table I. The size, distribution, and mean sequence length of clusters significantly vary for different motifs such as 'CG,' 'ATGC,' 'TATAAA,' 'GAATTC,' 'CACACACACA,' and 'AGGAGG.'.

1) 'ATGC' Sequence: The 'ATGC' motif demonstrates significant variability across bacterial clusters. Cluster 4 exhibits the highest mean occurrence (827.76) and the largest standard deviation (1343.67), with a maximum count reaching 21522. This data suggests a substantial enrichment of 'ATGC' in Cluster 4, likely associated with important biological activities such as replication or coding regions. Clusters 8 and 0 also show relatively high mean values (609.33 and 135.51, respectively), although with lower variability compared to Cluster 4. Clusters 1 and 7, on the other hand, have almost no "ATGC" events, with means below 2.0. The

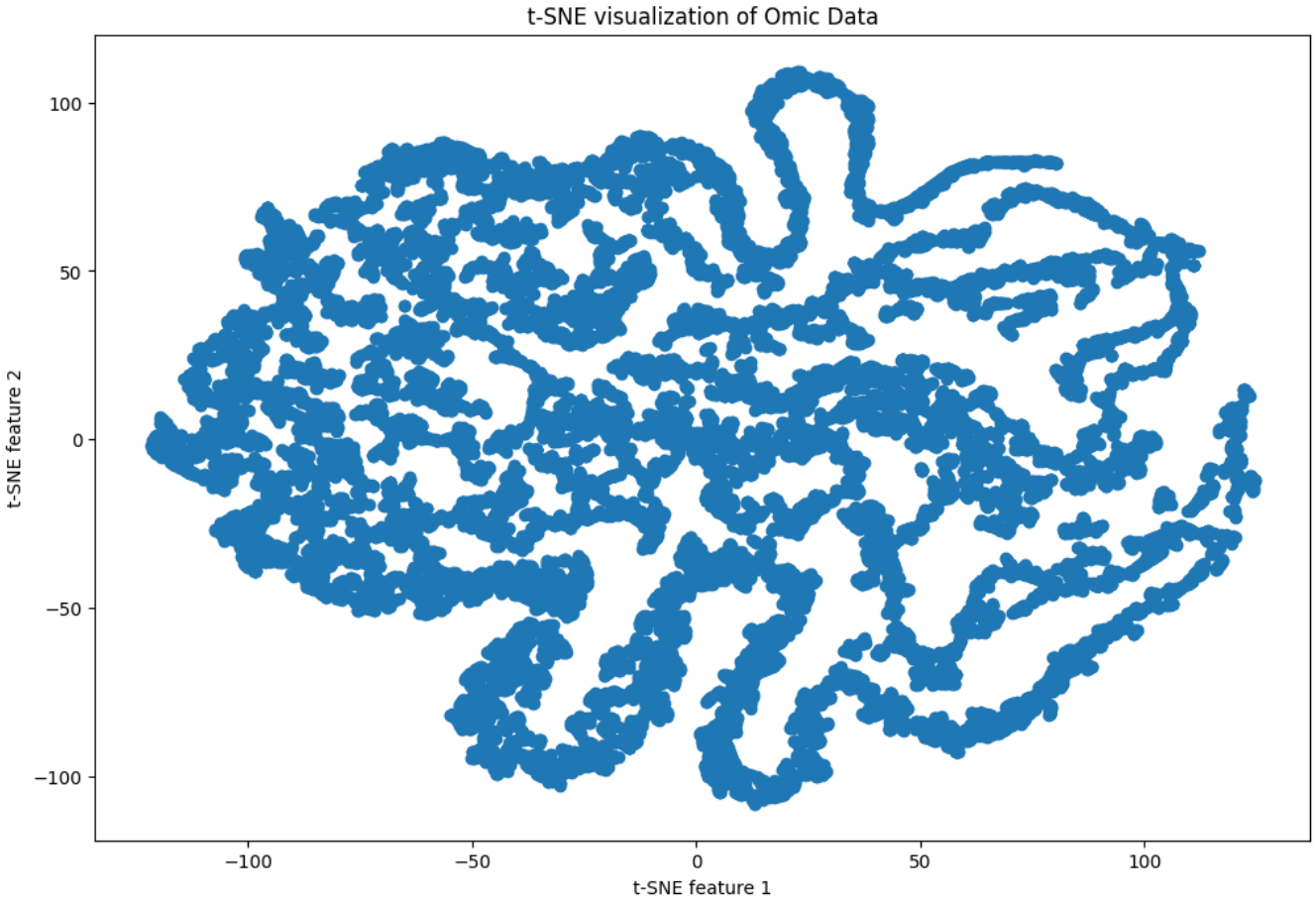

Fig. 6. Applying t-SNE (Self Made)

result suggests that this motif doesn't play a big role in these clusters, as shown in Table I.

2) 'TATAAA' Sequence: The 'TATAAA' motif, known for its role in transcription regulation, displays notable variability. Cluster 4 has the highest mean (23.33) and the largest variability, with a maximum count of 1430. The amounts of "TATAAA" in Clusters 0, 3, and 6 are moderate, with means ranging from 6.67 to 33.85. This means that transcription activity is moderate in these clusters. On the other hand, Clusters 1 and 7 don't show this motif very often (means less than 0.3), which could mean that transcription regulation for these clusters is weaker Table II.

3) 'CACACACACA' Sequence: The 'CACACACACA' motif is uncommon across the clusters, with Cluster 4 exhibiting the highest average occurrence (0.0166) and the greatest variability. Clusters 0 and 8 also show slightly elevated mean values of 0.0085 and 0.0505, respectively. In contrast, clusters 1, 2, 5, 6, and 7 have mean values close to zero, indicating that the CA5 motif likely does not serve a significant functional role in these clusters. Table III illustrates that its presence in certain clusters implies a potential connection to genomic stability.

4) 'AGGAGG' Sequence: The 'AGGAGG' motif is most common in Cluster 4, which has the highest mean (31.03) and variability, with a highest count of 538. Near bacterial ribosome binding sites (Shine-Dalgarno sequence), this motif frequently appears. Cluster 8 also shows a high mean (18.31) and variability, suggesting significant translational activity in these clusters. Clusters 0, 3, and 5 have average occurrences of 3.31 to 4.02, while Clusters 1, 2, and 7 have very low occurrences. This suggests that "AGGAGG" may not have a major impact on their translational regulation, as shown in Table IV.

5) 'GAATTC' Sequence: A big mean of 37.29 and a huge standard deviation of 59.97 show that Cluster 4 has the most of the "GAATTC" motif, which is a recognition site for the restriction enzyme EcoRI. The highest count was 816. Clusters 8, 0, and 3 also show moderate occurrences, with mean values of 37.97, 8.23, and 3.43, respectively. The fact that the mean is less than 0.2 suggests that "GAATTC" doesn't show up very often in Clusters 1, 2, and 7. This means that it may not play as important of a role in these clusters. This motif may be involved in genomic defense mechanisms in those clusters where it is enriched, as shown in Table V.

6) CG Sequence: The CG motif had the highest variability across clusters. Cluster 4 shows an extremely high mean of 19634.87, with a maximum count of 405576. Cluster 3 (mean: 1099.64) and Cluster 8 (mean: 6528.77) also have a lot of CG motifs. These may have something to do with CpG islands and epigenetic regulation. Clusters 1, 2, and 7 returned low mean values below 40, suggesting the motif's minimal importance in these genomes. In some clusters, Table VI shows that there are more CG motifs. This could mean that gene regulation or epigenetic activity is stronger.

TABLE I Data Statistics of 'ATGC' (Self Made)

Cluster	count	mean	std	min	25%	50%	75%	max
0	2351.0	135.510846	72.464687	13.0	82.00	128.0 1	83.0	478.0
1	2912.0	1.670330	2.071176	0.0	0.00	1.0	2.0	58.0
2	2248.0	2.977313	2.278735	0.0	1.00	3.0	4.0	13.0
3	3022.0	78.563534	47.473093	7.0	39.00	67.0	110.0	230.0
4	2292.0	827.755672	1343.667400	68.0	285.75	588.5	937.0	21522.0
5	2927.0	22.404510	26.727001	0.0	5.00	12.0	29.0	172.0
6	2892.0	22.772130	20.480824	0.0	8.00	15.0	28.0	145.0
7	2504.0	1.200879	1.209887	0.0	0.00	1.0	2.0	8.0
8	2473.0	609.325516	740.785644	123.0	292.00	395.0	597.0	15508.0
9	2749.0	9.903601	7.741960	0.0	4.00	8.0	14.0	119.0

TABLE II Data Statistics of 'TATAAA' (Self Made)

Cluster	count	mean	std	min	25%	50%	75%	max
0	2351.0	33.848150	21.025241	1.0	20.0	30.0	42.0	226.0
1	2912.0	0.021978	0.148961	0.0	0.0	0.0	0.0	2.0
2	2248.0	1.150356	1.342069	0.0	0.0	1.0	2.0	8.0
3	3022.0	6.668432	5.200889	0.0	3.0	6.0	10.0	35.0
4	2292.0	23.328970	72.962359	0.0	1.0	5.0	33.0	1430.0
5	2927.0	0.051589	0.256967	0.0	0.0	0.0	0.0	4.0
6	2892.0	7.906985	6.886790	0.0	3.0	6.0	11.0	50.0
7	2504.0	0.241214	0.539420	0.0	0.0	0.0	0.0	4.0
8	2473.0	106.212697	188.645418	2.0	20.0	51.0	119.0	4051.0
9	2749.0	0.660604	1.143990	0.0	0.0	0.0	1.0	12.0

TABLE III Data Statistics of 'CACACACACA' (Self Made)

Cluster	count	mean	std	min	25%	50%	75%	max
0	2351.0	0.008507	0.096381	0.0	0.0	0.0	0.0	2.0
1	2912.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
2	2248.0	0.001335	0.036515	0.0	0.0	0.0	0.0	1.0
3	3022.0	0.003971	0.062900	0.0	0.0	0.0	0.0	1.0
4	2292.0	0.016579	0.131090	0.0	0.0	0.0	0.0	2.0
5	2927.0	0.002050	0.045237	0.0	0.0	0.0	0.0	1.0
6	2892.0	0.000692	0.026293	0.0	0.0	0.0	0.0	1.0
7	2504.0	0.003195	0.126375	0.0	0.0	0.0	0.0	6.0
8	2473.0	0.050546	0.276274	0.0	0.0	0.0	0.0	4.0
9	2749.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0

TABLE IV Data Statistics of 'AGGAGG' (Self Made)

Cluster	count	mean	std	min	25%	50%	75%	max
0	2351.0	4.017439	3.955512	0.0	1.0	3.0	6.0	32.0
1	2912.0	0.067651	0.272193	0.0	0.0	0.0	0.0	2.0
2	2248.0	0.112100	0.351582	0.0	0.0	0.0	0.0	3.0
3	3022.0	2.089676	2.229671	0.0	1.0	2.0	3.0	19.0
4	2292.0	31.029232	49.702796	0.0	10.0	17.0	31.0	538.0
5	2927.0	3.314315	4.390882	0.0	0.0	2.0	5.0	35.0
6	2892.0	0.703320	1.201787	0.0	0.0	0.0	1.0	20.0
7	2504.0	0.110623	0.501837	0.0	0.0	0.0	0.0	10.0
8	2473.0	18.305297	37.835738	0.0	5.0	8.0	18.0	1098.0
9	2749.0	0.425973	0.856983	0.0	0.0	0.0	1.0	17.0

TABLE V Data Statistics of 'GAATTC' (Self Made)

Cluster	count	mean	std	min	25%	50%	75%	max
0	2351.0	8.233092	5.707265	0.0	4.0	7.0	11.0	35.0
1	2912.0	0.055632	0.252087	0.0	0.0	0.0	0.0	3.0
2	2248.0	0.198843	0.468936	0.0	0.0	0.0	0.0	3.0
3	3022.0	3.432826	3.001371	0.0	1.0	3.0	5.0	20.0
4	2292.0	37.286649	59.973713	0.0	14.0	24.0	40.0	816.0
5	2927.0	0.837718	1.737362	0.0	0.0	0.0	1.0	15.0
6	2892.0	1.694329	1.975795	0.0	0.0	1.0	2.0	13.0
7	2504.0	0.092252	0.323343	0.0	0.0	0.0	0.0	3.0
8	2473.0	37.970886	62.034258	1.0	9.0	18.0	37.0	1092.0
9	2749.0	0.441615	0.738788	0.0	0.0	0.0	1.0	5.0

```
plt.scatter(X_tsne[:, 0], X_tsne[:, 1], c=clusters, cmap='viridis')
```

```
<matplotlib.collections.PathCollection at 0x7c8220ec6050>
```

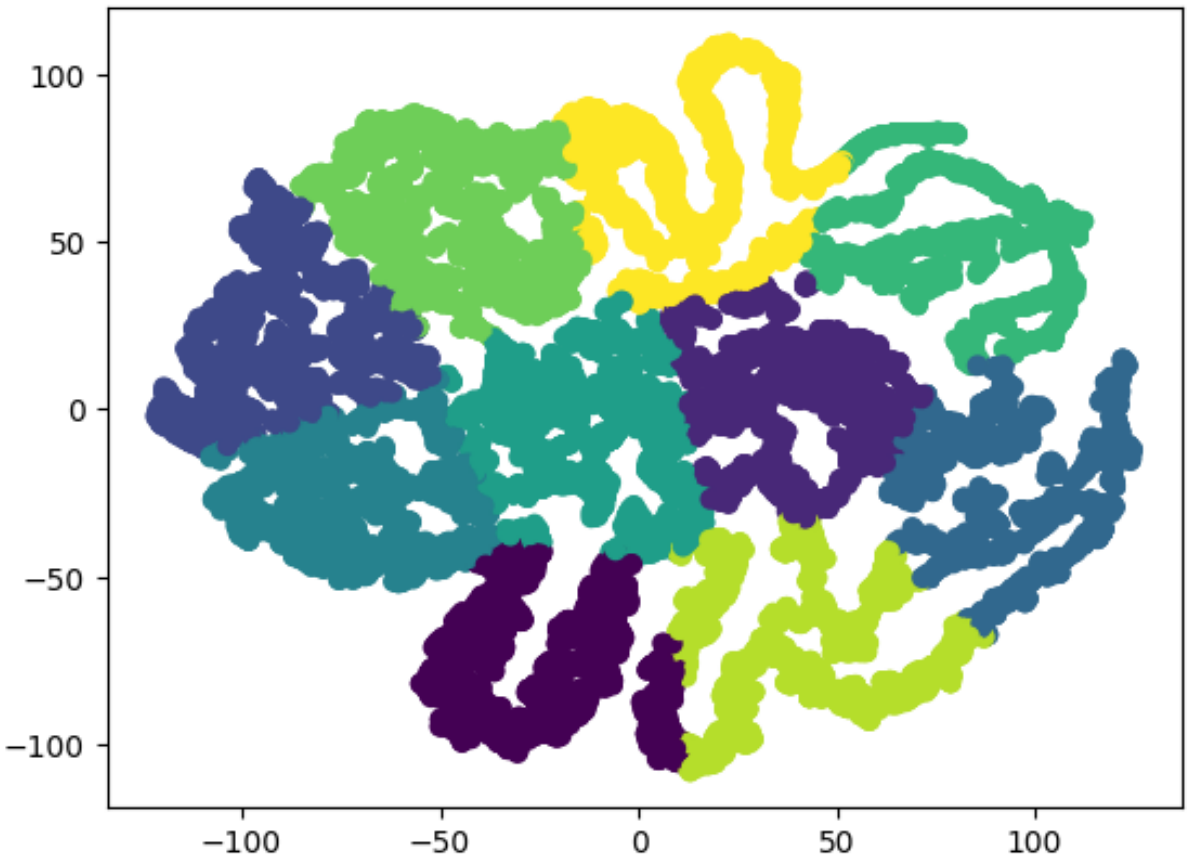

Fig. 7. Assign Label to t-SNE outcome (Self Made)

TABLE VI Data Statistics of 'CG' (Self Made)

Cluster	count	mean	std	min	25%	50%	75%	max
0	2351.0	930.387495	63.041905	30.0	488.5	868.0	1291.00	2850.0
1	2912.0	36.981113	21.918888	0.0	22.0	32.0	46.00	198.0
2	2248.0	22.008007	15.121697	0.0	11.0	19.0	30.00	90.0
3	3022.0	1099.642621	804.549008	88.0	453.0	815.5	1668.75	3804.0
4	2292.0	19634.874782	30296.551391	1798.0	7882.0	12516.0	20585.75	405576.0
5	2927.0	1081.102152	1124.866292	62.0	214.0	606.0	1618.50	5683.0
6	2892.0	139.123444	121.399613	5.0	53.0	93.0	172.00	629.0
7	2504.0	12.997604	8.677974	0.0	7.0	12.0	18.00	53.0
8	2473.0	6528.765467	8068.880748	379.0	3183.0	4653.0	7176.00	179596.0
9	2749.0	144.676610	113.654030	0.0	69.0	106.0	187.00	803.0

H. Nucleotide Frequency Calculation:

Looking at the differences in the number of nucleotides in different clusters shows that they have different makeups and could have different functions. In Cluster 0, adenine (A) and thymine (T) are much more common than cytosine (C) and guanine (G), showing that it is an AT-rich cluster. Cluster 0 might be made up of non-coding regions or places like promoter sequences that have a lot of AT-rich segments. Cluster 1 is GC-rich, with guanine (G) and cytosine (C) levels dominating. Usually, GC pairs are linked to coding sequences or genomic regions that need to be more stable because the hydrogen bonds between them are stronger. Cluster 2: This group has a balanced amount of AT and GC content, which shows areas with different functions, like sequences that do both coding and regulatory work. Minimal levels of ambiguous bases suggest high sequence quality. Cluster 3 has almost exactly the same number of all four nucleotides. This finding indicates that there are conserved areas in the genome that may be limited by evolution or play important structural or functional roles.

For example, Cluster 4 has a lot of GC, which means it probably has coding regions or structurally complex parts like CpG islands. The slight presence of ambiguous bases

indicates some variability or sequence uncertainty. Cluster 5: This cluster shows the strongest GC bias among all, with guanine (G) and cytosine (C) dominating. Such sequences may correspond to stable regions, possibly linked to specific gene expression or protein-coding regions. Cluster 6: Similar to Cluster 0, this cluster has a high AT content, pointing to regions that are regulatory in nature or less thermodynamically stable. The low occurrence of ambiguous bases suggests good data quality. Cluster 7: This cluster is moderately balanced but leans toward being AT-rich. It may include sequences with regulatory or gene-associated roles requiring moderate stability. Cluster 8 demonstrates a balance in nucleotide composition with a slight tilt toward AT-richness. Such sequences may indicate regions with mixed functional elements like a combination of coding and regulatory segments. Cluster 9: The nucleotides in this cluster are evenly spread out, which suggests that they are made up of conserved sequences that may play important structural or functional roles in the genome, as shown in Fig. 8.

III. Results

In this section of the results, we have utilized six different motif sequences, namely 'TATAAA,' 'ATGC,' 'CACACACACA,' 'GAATTC,' 'AGGAGG,' and 'CG.' We meticulously examine each motif in our bacterial omics data. The first step is to use the statistical non-parametric Kruskal-Wallis H-Test to look for differences in frequency between clusters for each motif sequence. Next, the pairwise cluster comparisons that use the Bonferroni Correction and the Mann-Whitney U Test are a good way to control the type I error in multiple comparisons. Finally, we also plot the heat map of different motif sequences.

A. Motif 'ATGC'

The Kruskal-Wallis H test was used to see if there were important differences in how the ATGC values were spread out across 10 groups (Cluster 0 through Cluster 9). This test is a non-parametric statistical method used to compare distributions across multiple independent groups when the assumption of normality may not hold. The test, which used the kruskal() function from the scipy.stats module, gave a Kruskal-Wallis H statistic of 22910.5181 and a p-value of 0.0, which means the result was very significant, as shown in Fig. 9. The null hypothesis for this test assumes that all clusters have the same distribution of ATGC values. However, the tiny p-value (essentially zero) led to the rejection of the null hypothesis. This result confirms that there are significant differences in the distribution of ATGC values between the clusters. These findings suggest that the clusters are not homogeneous with respect to the presence or frequency of ATGC motifs. Such significant differences may point to distinct nucleotide sequence patterns or motif distributions across the clusters. These results highlight the effectiveness of the clustering methodology in capturing meaningful variations in the dataset, warranting further analysis to understand the biological or computational implications of these differences.

1) Statistical Significance of Motif Distribution Across Clusters 'ATGC': The results of the Mann-Whitney U test and Bonferroni correction indicate that the "ATGC" motif exhibits a distinct distribution within the bacterial genome clusters. This finding suggests that this motif has different functions in different parts of the genome. For instance, the comparison between Cluster 9 and Cluster 2 yielded a U statistic of 5,170,620.5 with a p-value of 0.0, indicating a highly significant difference. The finding suggests that Cluster 9, with its balanced nucleotide composition, may be made up of areas with functional roles that have been conserved, while Cluster 2, with its AT-rich sequences, may be made up of regulatory or promoter regions. As you can see in Table VII(a), the comparison between Cluster 8 and Cluster 6 (U = 7,151,912.0, p = 0.0) shows that these two groups have different genomic features that are probably the result of different evolutionary pressures or functional needs. The Mann-Whitney U test indicates that the ATGC motif is spread out in many different ways across clusters, which supports its role in many different genomic functions.

2) Pairwise Cluster Comparison Using Mann-Whitney U Test for 'ATGC': After using the Bonferroni correction to account for multiple comparisons, we kept the importance of the differences in the distribution of the "ATGC" motif across all clusters. There was a big difference in the numbers between Cluster 9 and Cluster 6 (U = 2,236,214.0, Bonferroni-adjusted p = 3.28×10^{-178}), which supported the idea that these clusters have different functions. Furthermore, the big difference between Cluster 5 and Cluster 4 (U = 9,419.5, Bonferroni adjusted p value = 0.0) suggests that the "ATGC" motif may have different roles, such as coding functions in Cluster 5 and regulatory functions in Cluster 4 (Table VII(b)). The results demonstrate the diverse distribution of the "ATGC" motif across various bacterial groups. This is likely because the bacteria have adapted to meet different functional or environmental needs, as shown in Fig. 21.

B. Motif 'TATAAA'

The Kruskal-Wallis H-Test is then used to see how the "TATAAA" motif, which is made up of thymine (T) and adenine (A) nucleotides, is spread out among the groups of bacteria. This motif, commonly known as the TATA box, is a critical component of promoter regions and plays a significant role in transcription initiation. The analysis produced a Kruskal-Wallis statistic of 20149.857033237466, and the p-value was effectively zero (p_value: 0.0), as shown in Fig. 10. The very low p-value, which is 0.05, means that the null hypothesis is not true. This indicates that the "TATAAA" motif distribution is completely unique between the clusters. After the Kruskal-Wallis test, the next step is to apply the Mann-Whitney U statistic.

1) Statistical Significance of Motif Distribution Across Clusters 'TATAAA': The Mann-Whitney U test indicated that the frequency distribution of the "TATAAA" motif, which is usually thought of as a key promoter element, was completely unique between clusters. As one of the most illustrative examples of such testing, Cluster 9 vs. Cluster 2 resulted in U = 2,333,550.5, p = 1.02×10^{-59}. Cluster 9 has a more even distribution of nucleotides, which suggests that it may have a wider range of functions. On the other hand, Cluster 2's high AT content suggests that it is more likely connected with controlling transcription. There were also significant differences between Cluster 9

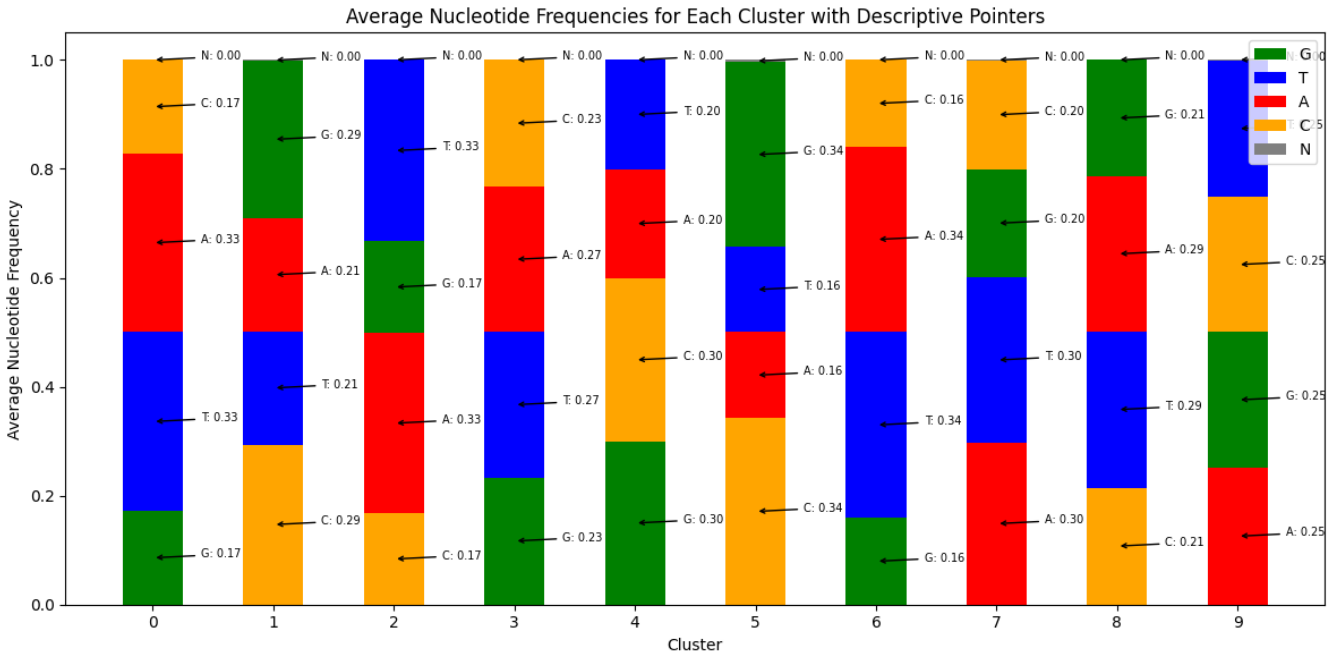

Fig. 8. Nucleotide Frequency of Motif(Self Made)

```
[ ]  from scipy.stats import kruskal

     # Perform the Kruskal-Wallis H test
     kruskal_statistic, kruskal_p_value = kruskal(
         data[data['Cluster'] == 0]['ATGC_Found'],
         data[data['Cluster'] == 1]['ATGC_Found'],
         data[data['Cluster'] == 2]['ATGC_Found'],
         data[data['Cluster'] == 3]['ATGC_Found'],
         data[data['Cluster'] == 4]['ATGC_Found'],
         data[data['Cluster'] == 5]['ATGC_Found'],
         data[data['Cluster'] == 6]['ATGC_Found'],
         data[data['Cluster'] == 7]['ATGC_Found'],
         data[data['Cluster'] == 8]['ATGC_Found'],
         data[data['Cluster'] == 9]['ATGC_Found']
     )

     print(f"Kruskal-Wallis H Test Statistic: {kruskal_statistic}, P-value: {kruskal_p_value}")
     if kruskal_p_value > 0.05:
         print('Fail to reject the null hypothesis - no significant difference between clusters')
     else:
         print('Reject the null hypothesis - significant differences between clusters')

     Kruskal-Wallis H Test Statistic: 22910.51811487161, P-value: 0.0
     Reject the null hypothesis - significant differences between clusters
```

Fig. 9. Kruskal-Wallis H-Test 'ATGC' (Self Made)

and Cluster 6 (U = 484,948.0, p = 0.0), which suggests that these clusters are controlled by different systems, as shown in Table VIII(a). The "TATAAA" motif plays a big part in controlling transcription, and these results indicate that it can be found in different parts of the genome.

2) Pairwise Cluster Comparison Using Mann-Whitney U Test for 'TATAAA': Using the Bonferroni correction on the "TATAAA" motif backs up the results from the Mann-Whitney U test even more when it comes to the significant differences. For instance, comparing Cluster 9 to Cluster 7 (U = 4,116,330.5, Bonferroni adjusted p value = 9.03×10^{-54}) showed that these two groups have completely unique promoter profiles, which is another sign that they have different transcriptional needs. However, there was no statistical dif-

```
[ ]  from scipy.stats import kruskal

     # Perform the Kruskal-Wallis H test
     kruskal_statistic, kruskal_p_value = kruskal(
         data[data['Cluster'] == 0]['TATAAA_Found'],
         data[data['Cluster'] == 1]['TATAAA_Found'],
         data[data['Cluster'] == 2]['TATAAA_Found'],
         data[data['Cluster'] == 3]['TATAAA_Found'],
         data[data['Cluster'] == 4]['TATAAA_Found'],
         data[data['Cluster'] == 5]['TATAAA_Found'],
         data[data['Cluster'] == 6]['TATAAA_Found'],
         data[data['Cluster'] == 7]['TATAAA_Found'],
         data[data['Cluster'] == 8]['TATAAA_Found'],
         data[data['Cluster'] == 9]['TATAAA_Found']
     )

     print(f"Kruskal-Wallis H Test Statistic: {kruskal_statistic}, P-value: {kruskal_p_value}")
     if kruskal_p_value > 0.05:
         print('Fail to reject the null hypothesis - no significant difference between clusters')
     else:
         print('Reject the null hypothesis - significant differences between clusters')

     Kruskal-Wallis H Test Statistic: 20149.857033237466, P-value: 0.0
     Reject the null hypothesis - significant differences between clusters
```

Fig. 10. Kruskal-Wallis H-Test for 'TATAAA' (Self Made)

ference between Cluster 6 and Cluster 4 (U = 3,359,474.5, Bonferroni adjusted p-value = 0.39); this could mean that both clusters are using similar ways to start transcription, as shown in Table VIII(b). In fact, these results indicate that the pattern of distribution for the "TATAAA" motif has been specifically changed to meet certain functional and regulatory needs, with most of the clusters being more different and only a few being similar, as shown in Fig. 22.

C. Motif 'CACACACACA'

We will next analyze a specific sequence known as "CACACACACA." Our bacterial clusters contain five cytosine-adenine repeats that make up this sequence. We employ the Kruskal-Wallis H-Test, as shown in Fig. 11, to analyze the data. This statistical test is practical when we don't have any assumptions about the distribution of our data. The test produced a statistic of 483.7493979046291 and a p-value of $1.7313265380364424 \times e^{-98}$, which is relatively small as shown in Fig. 11. These results indicate significant differences in the 'CA5' motif distribution across the clusters. Therefore, we can reject the null hypothesis, which suggests that the clusters have distinct 'CA5' motif frequencies. Based on these findings, we need to explore the functional implications of this motif in bacterial genomics to better understand its role.

1) Statistical Significance of Motif Distribution Across Clusters 'CACACACACA': We used the Mann-Whitney U test and Bonferroni correction to determine which groups of the motif CA5 were statistically significant. The motif CA5 is made up of five cytosine-adenine repeats. Pairwise comparisons between the clusters illustrate several different levels of statistical significance. Taking the example of Cluster 0 vs. Cluster 8, the U statistic is 2820944.5, with a p-value of 8.84×10^{-12} . The result indicates that there was a highly statistically significant difference in the distribution of this motif between these two clusters. In the same way, comparing Cluster 4 to Cluster 2 gave us a U statistic of 2538056.5 and a p-value of 9.4×10^{-8}, which shows that the motif's appearance in these clusters is very different. However, not all pairs show this, with examples like Cluster 2 and Cluster 7 providing a U statistic of 2816001.0 with a p-value of 0.57, as shown in Table IX(a). which indicates a similar distribution in the distribution of this motif, thus indicating similar distributions. These findings show that the CA5 motif is not the same in all bacterial clusters. This may indicate that the genome is organized differently or that regulatory mechanisms differ.

2) Pairwise Cluster Comparison Using Mann-Whitney U Test for 'CA5': Adding the Bonferroni correction to the p-values from the Mann-Whitney U test made a number of comparisons more statistically significant. For example, the comparison between Cluster 8 and Cluster 0 (corrected p-value: True) reinforces the previously identified significant difference. Other examples include Cluster 4 versus Cluster 9 (corrected p-value: True) and Cluster 8 versus Cluster 6 (corrected p-value: True), confirming significantly different distributions of the CA5 motif. Other comparisons, like Cluster 2 vs. Cluster 6 (corrected p-value: False), are no longer significant after correction, which shows how conservative the Bonferroni adjustment is. As shown in Table IX(b) and Fig. 23, these results show how strong certain significant differences are and how important it is to be careful when

```python
from scipy.stats import kruskal

# Perform the Kruskal-Wallis H test
kruskal_statistic, kruskal_p_value = kruskal(
    data[data['Cluster'] == 0]['CA5_Found'],
    data[data['Cluster'] == 1]['CA5_Found'],
    data[data['Cluster'] == 2]['CA5_Found'],
    data[data['Cluster'] == 3]['CA5_Found'],
    data[data['Cluster'] == 4]['CA5_Found'],
    data[data['Cluster'] == 5]['CA5_Found'],
    data[data['Cluster'] == 6]['CA5_Found'],
    data[data['Cluster'] == 7]['CA5_Found'],
    data[data['Cluster'] == 8]['CA5_Found'],
    data[data['Cluster'] == 9]['CA5_Found']
)

print(f"Kruskal-Wallis H Test Statistic: {kruskal_statistic}, P-value: {kruskal_p_value}")
if p > 0.05:
    print('Fail to reject the null hypothesis - no significant difference between clusters')
else:
    print('Reject the null hypothesis - significant differences between clusters')
```

```
Kruskal-Wallis H Test Statistic: 483.7493979046291, P-value: 1.7313265380364424e-98
Reject the null hypothesis - significant differences between clusters
```

Fig. 11. Kruskal-Wallis H-Test for 'CACACACACA' (Self Made)

interpreting borderline cases.

D. Motif 'AGGAGG'

We conducted the Kruskal-Wallis H test to determine if the presence of 'AGGAGG' varied significantly across different bacterial clusters. The test resulted in a high H statistic of 17334.04 and a p-value of 0.0, indicating strong statistical significance, as shown in Fig. 12. Since the p-value is below the standard significance threshold of 0.05, the null hypothesis, which assumes no significant differences between clusters, is rejected. This finding suggests that the presence of "AGGAGG" varies a lot between bacterial clusters, which means that different groups have different genes. Such variations may be linked to differences in bacterial adaptation, mutation rates, or evolutionary processes. The results highlight the importance of genetic diversity in bacterial populations and suggest that certain clusters may have unique genetic characteristics that influence their survival, pathogenicity, or environmental adaptability.

1) Statistical Significance of Motif Distribution Across Clusters' AGGAGG': In the same way, we looked at the pattern "AGGAGG," which is usually found near bacterial ribosome-binding sites, to see how it was spread out among the clusters. Most of the comparisons showed significant deviations; for instance, Cluster 0 versus Cluster 8 resulted in a U statistic of 1152672.5 and a p-value of 1.19×10^{-289} , highlighting a big difference. Similarly, Cluster 9 versus Cluster 2 gave a U statistic of 2436011.5 with a p-value of 1.47×10^{-72} , thus confirming extreme variability. Some comparisons, like Cluster 2 vs. Cluster 7, did not show statistical significance, with a U statistic of 2852862.5 and a p-value of 0.11. This means that there is a uniform distribution in that case, as shown in Table X(a). These findings show that the 'AGGAGG' motif can have different functions in different bacterial genomes. They may also show that it has a different role in starting translation in different clusters.

2) Pairwise Cluster Comparison Using Mann-Whitney U Test for 'AGGAGG': Most of these comparisons remained robust after the application of the Bonferroni correction. For example, Cluster 0 versus Cluster 8, which had an initial p-value of 1.19×10^{-289} , remained significant after correction, with a corrected p-value of True. Likewise, Cluster 9 versus Cluster 4 remained significant, with a corrected p-value of True underscoring different motif distributions. On the other hand, some comparisons, such as Cluster 2 versus Cluster 7, became non-significant after correction, with a corrected p-value of False. Table X(b) and Fig. 24 demonstrate the stringency of this correction method. These results back up the reliability of significant results and show that the 'AGGAGG' motif is spread out in a complex way among the bacterial clusters.

E. Motif 'GAATTC'

Next, we examine 'GAATTC', the EcoRI recognition site. Before we look at these sequences, we use the Kruskal-Wallis H-Test to see how common this motif is among bacterial groups. The test showed significant differences in the frequency of this motif among the clusters, with a statistic of 19736.04433017053 and a p-value of 0.0. The result indicates that the distribution of the 'GAATTC' motif is not

```python
from scipy.stats import kruskal

# Perform the Kruskal-Wallis H test for 'AGGAGG' across all clusters
kruskal_statistic, kruskal_p_value = kruskal(
    data[data['Cluster'] == 0]['AGGAGG_Found'],
    data[data['Cluster'] == 1]['AGGAGG_Found'],
    data[data['Cluster'] == 2]['AGGAGG_Found'],
    data[data['Cluster'] == 3]['AGGAGG_Found'],
    data[data['Cluster'] == 4]['AGGAGG_Found'],
    data[data['Cluster'] == 5]['AGGAGG_Found'],
    data[data['Cluster'] == 6]['AGGAGG_Found'],
    data[data['Cluster'] == 7]['AGGAGG_Found'],
    data[data['Cluster'] == 8]['AGGAGG_Found'],
    data[data['Cluster'] == 9]['AGGAGG_Found']
)

print(f"Kruskal-Wallis H Test Statistic for AGGAGG: {kruskal_statistic}, P-value: {kruskal_p_value}")
if p > 0.05:
    print('Fail to reject the null hypothesis - no significant difference between clusters')
else:
    print('Reject the null hypothesis - significant differences between clusters')
```

```
Kruskal-Wallis H Test Statistic for AGGAGG: 17334.042008495857, P-value: 0.0
Reject the null hypothesis - significant differences between clusters
```

Fig. 12. Kruskal-Wallis H-Test for 'AGGAGG' (Self Made)

uniform. Since the p-value was below the standard significance threshold of p= 0.05 across all cluster comparisons, it suggests potential variations in restriction sites. These significant results have implications for further research, as shown in Fig. 13.

1) Statistical Significance of Motif Distribution Across Clusters 'GAATTC': This motif, 'GAATTC,' represents the recognition site of the EcoRI restriction enzyme. We used the Mann-Whitney U test to compare this motif across bacteria clusters. In most of the paired comparisons, the results showed statistically significant differences in the distribution of the motif. For example, the U statistic for Cluster 0 versus Cluster 3 is 5540881.5, with a p-value of 2.07×10^{-274} . We may, therefore, say that Cluster 0 has a significantly different generation of motifs compared to Cluster 3. The U statistic of Cluster 2 compared to Cluster 9 is 2588588.5, with a corresponding p-value of 2.71×10^{-38}, thereby strongly confirming observed variability in motif distribution. Some comparisons showed more interesting results, like the one between Cluster 3 and Cluster 7, which had a U statistic of 6996482.0 and a p-value of 0.0. This again showed that the patterns in the motifs between the clusters were different. Based on this, our findings show that the 'GAATTC' motif has different functions or structures in the genome, as shown in Table XI(a).

2) Pairwise Cluster Comparison Using Mann-Whitney U Test for 'GAATTC': The Bonferroni correction for multiple testing made sure that the differences that were statistically significant between the pairs were real and strong. This lowers the chance of getting false positives.

For instance, the comparison of Cluster 4 and Cluster 8 produced a corrected p-value of True, showing a significant difference in motif distribution. The result indicates that these clusters are unique in their properties, perhaps reflecting underlying biological heterogeneity in motif frequency or function. Similarly, Cluster 6 and Cluster 9 showed a significant difference, with a corrected p-value of True, as revealed in Table XI(b).

These results show that the "GAATTC" motif, which is usually found next to the EcoRI recognition site, might have a functional role. Bacterial genomes usually keep these spots because they are important for restriction-modification systems and might play a part in how organisms have changed over time. The results, shown in Fig. 25, show that the statistical analysis is strong and reliable. They also point to a possible adaptive role for the "GAATTC" motif in different bacterial clusters.

F. Motif 'CG'

We used the Kruskal-Wallis H-Test to look at how often the pattern 'CG' dinucleotides showed up in the motif 'CG' sequence in our bacterial data. The Kruskal-Wallis statistic we got from our analysis was 23,464.1269, and the p-value was 0.0. Fig. 14 demonstrates a significant difference in the motif distribution across the clusters. The 'CG' pattern is often associated with CpG islands, which play a crucial role in gene regulation through methylation. We observed that the distribution of this motif was not uniform across all clusters. This critical finding suggests that different groups of bacteria have different methylation potentials and levels of gene regulatory activity.

1) Statistical Significance of Motif Distribution Across Clusters 'CG': The motif CG has often been associated with CpG islands and methylation sites, and hence a Mann-Whitney U test was conducted to get a deeper understanding

```python
from scipy.stats import kruskal

# Perform the Kruskal-Wallis H test for 'GAATTC' across all clusters
kruskal_statistic, kruskal_p_value = kruskal(
    data[data['Cluster'] == 0]['GAATTC_Found'],
    data[data['Cluster'] == 1]['GAATTC_Found'],
    data[data['Cluster'] == 2]['GAATTC_Found'],
    data[data['Cluster'] == 3]['GAATTC_Found'],
    data[data['Cluster'] == 4]['GAATTC_Found'],
    data[data['Cluster'] == 5]['GAATTC_Found'],
    data[data['Cluster'] == 6]['GAATTC_Found'],
    data[data['Cluster'] == 7]['GAATTC_Found'],
    data[data['Cluster'] == 8]['GAATTC_Found'],
    data[data['Cluster'] == 9]['GAATTC_Found']
)

print(f"Kruskal-Wallis H Test Statistic for GAATTC: {kruskal_statistic}, P-value: {kruskal_p_value}")
if p > 0.05:
    print('Fail to reject the null hypothesis - no significant difference between clusters')
else:
    print('Reject the null hypothesis - significant differences between clusters')
```

```
Kruskal-Wallis H Test Statistic for GAATTC: 19736.04433017053, P-value: 0.0
Reject the null hypothesis - significant differences between clusters
```

Fig. 13. Kruskal-Wallis H-Test for 'GAATTC' (Self Made)

```python
# Perform the Kruskal-Wallis H test
kruskal_statistic, kruskal_p_value = kruskal(
    data[data['Cluster'] == 0]['CG_Found'],
    data[data['Cluster'] == 1]['CG_Found'],
    data[data['Cluster'] == 2]['CG_Found'],
    data[data['Cluster'] == 3]['CG_Found'],
    data[data['Cluster'] == 4]['CG_Found'],
    data[data['Cluster'] == 5]['CG_Found'],
    data[data['Cluster'] == 6]['CG_Found'],
    data[data['Cluster'] == 7]['CG_Found'],
    data[data['Cluster'] == 8]['CG_Found'],
    data[data['Cluster'] == 9]['CG_Found']
)

print(f"Kruskal-Wallis H Test Statistic: {kruskal_statistic}, P-value: {kruskal_p_value}")
if p > 0.05:
    print('Fail to reject the null hypothesis - no significant difference between clusters')
else:
    print('Reject the null hypothesis - significant differences between clusters')
```

```
Kruskal-Wallis H Test Statistic: 23464.126900645257, P-value: 0.0
Reject the null hypothesis - significant differences between clusters
```

Fig. 14. Kruskal-Wallis H-Test for 'CG' (Self Made)

of its distribution across the clusters. Substantial differences were indicated in the occurrence of this motif by the test: for instance, Cluster 0 versus Cluster 5 gave a U statistic of 3797622.5, with a p-value of 8.74×10^{-11} , showing highly significant variability in motif distribution. The comparison between Cluster 2 and Cluster 7 produced a U statistic of 3856358.0 and a p-value of 5.38×10^{-108}, which means that there are important differences between the two groups, as shown in Table XII(a). There is a chance that the "CG" motif has different regulatory or structural genomic functions, as

most comparisons show.

2) Pairwise Cluster Comparison Using Mann-Whitney U Test for 'CG': The statistically significant results in the Mann-Whitney U test were mostly the same after the Bonferroni correction was applied. This gave the original results more weight. The change in the significance level to account for multiple comparisons showed that the differences between the pairs of clusters were important.

When we changed the p-value for the comparison between Cluster 3 and Cluster 5, we saw that there was still a statistically significant difference in the motif distributions between the two groups. Furthermore, the statistical difference between Cluster 0 and Cluster 6 was still very big, showing that genomic motif patterns are consistently different. Also, the comparison between Cluster 1 and Cluster 4 produced a p-value that stayed statistically significant even after correction. This indicates that the two groups have different patterns of motif distribution.

All of these results indicate that the "CG" motif is important in biology, especially when it comes to telling important genomic differences between different clusters. Using the Bonferroni correction gave these results more credibility by removing any false positives. This showed that the differences seen were not just random but actually caused by differences in the genome. Table XII(b) and Fig. 26 show that this statistical strength suggests the "CG" motif may have an effect on the structure and control of genomics.

IV. Heat Map Result of Motif

This heatmap for the 'ATGC' motif probably points out clusters that have different levels of motif abundance. For instance, clusters such as 4 and 8 may contain higher intensities since their counts are high, while clusters like 1 and 7 will have lower intensities due to smaller numbers of motif occurrences. A lot of the same motifs in a single cluster could mean that they are involved in important biological processes or conserved regions, which could be connected to important genomic functions, like replication origins. However, low prevalence may indicate that these clusters are less important for regulation or have less sequence variation, as shown in Fig. 15.

When you look at heat maps of the "TATAAA" motif, clusters 0, 4, 8, etc. will stand out because they have a lot of motifs, while clusters 1 and 7 will have the least amount of intensity. The 'TATAAA' motif is generally associated with the promoter regions in bacterial genomes. As seen in Fig. 16, clusters with a lot of motifs may be genes that play a big role in controlling transcription, while clusters with few motifs may be non-coding regions or places where transcription isn't happening much.

For the CA5 motif, clusters 8 and 4 have far more motif intensity compared to other clusters, having either negligible or no presence. It seems like this finding suggests that the CA5 motif may be very important in these groups, possibly showing repetitive or structural patterns in the genome, as seen in Fig. 17. High motif intensity means that the area might be biologically important, while low-intensity clusters mean that the area is not structurally or functionally important.

Similarly, clusters 4, 8, and 5 represent high intensity, i.e., dense frequencies of motifs, while clusters 1, 2, and

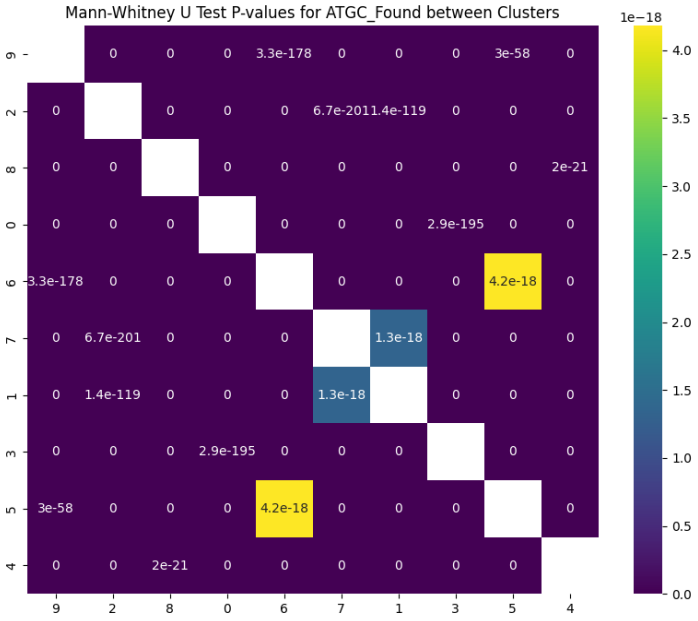

Fig. 15. Heat Map of 'ATGC'

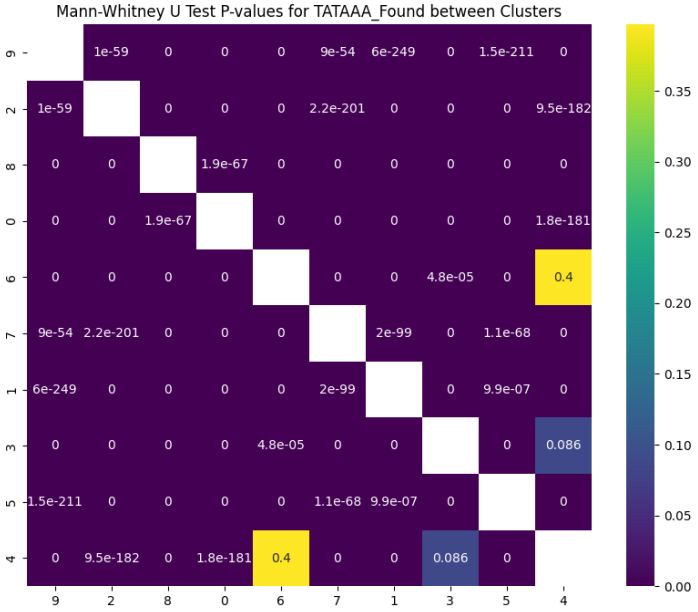

Fig. 16. Heat Map of 'TATAAA'

7 represent low intensity, i.e., lower frequencies of motifs. The observed trend shows that the clusters are not all the same in terms of how they work, which could be because of differences in genomic activity or structural features.

The 'AGGAGG' is a known Shine-Dalgarno motif, and it plays a role in translation initiation by enhancing ribosome binding. Areas that are actively translating may link to clusters with a high concentration of this motif, which could potentially serve as ribosome binding sites. According to Fig. 18, clusters 4 and 8 possess high motif intensity that can be ribosome binding regions that are actively translating. Clusters 1 and 7 are low intensity, which can be low translation activity regions or perhaps untranslated regions.

The "GAATTC" motif matches the EcoRI recognition site, which is an important part for restriction enzyme activity. As shown in Fig. 19, the motif frequency is high, indicating the

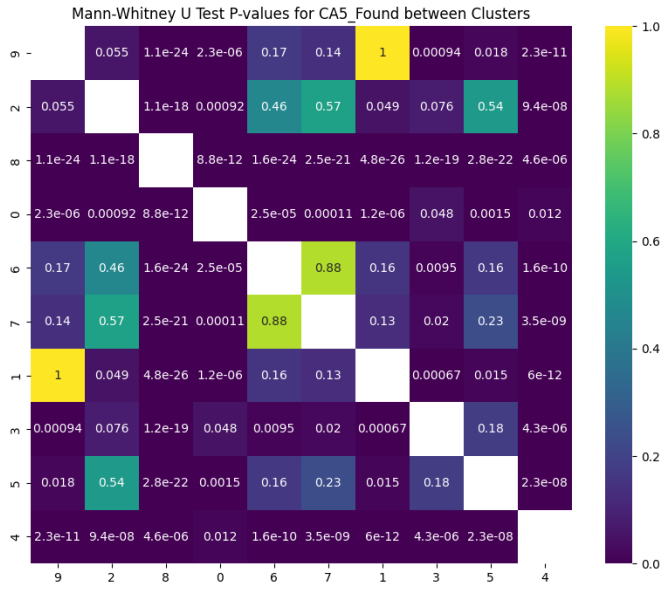

Fig. 17. Heat Map of 'CACACACACA'

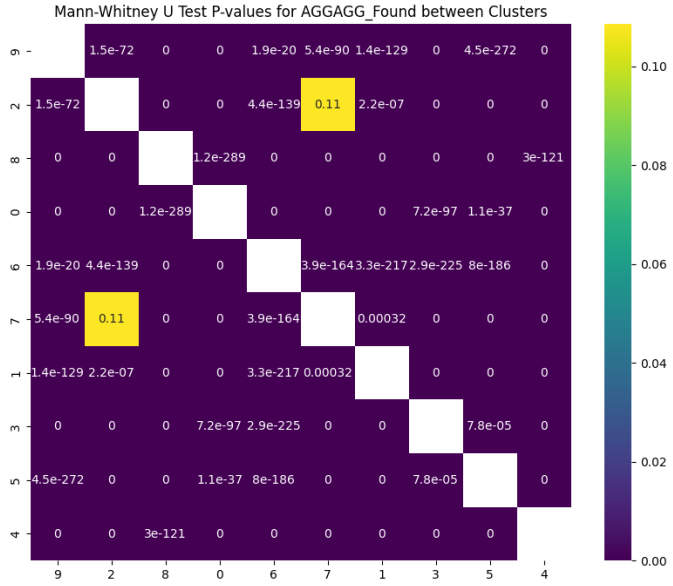

Fig. 18. Heat Map of 'AGGAGG'

presence of conserved cleavage sites. This suggests evolutionary pressures may preserve these sites because of their functional importance. Clusters 4, 8, and 5 have a lot of motif intensity, which means they might have important structural areas, such as ones that can be affected by horizontal gene transfer. Clusters 7 and 1 have low motif intensity, which means that these areas aren't affected as much by these kinds of evolutionary pressures.

The 'CG' motif is strongly associated with CpG islands, which play a crucial role in epigenetic regulation through DNA methylation. Fig. 20 shows that areas with more intense clusters, like clusters 4 and 8, may be places where epigenetic regulation or gene expression is active. The fact that these clusters contain CpG islands suggests that they might be regulatory hotspots, which could have big effects on gene activity. On the other hand, clusters with lower intensities might be regions with less regulatory influence, which could be non-coding sequences or regions that don't have much functional significance.

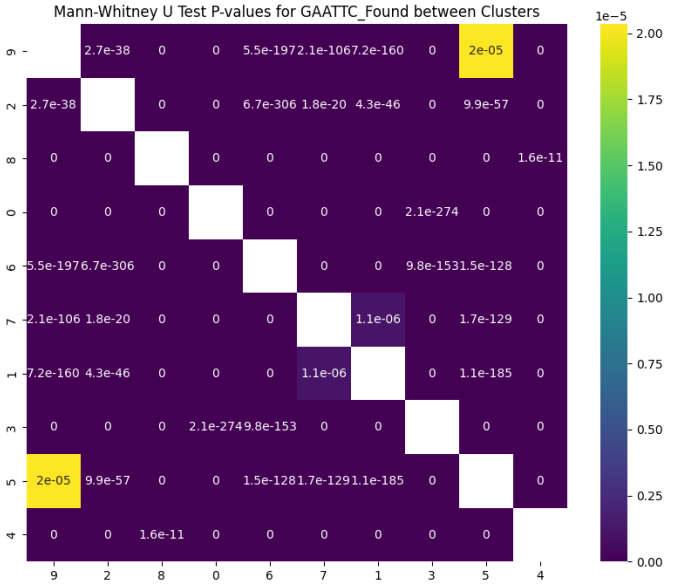

Fig. 19. Heat Map of 'GAATTC'

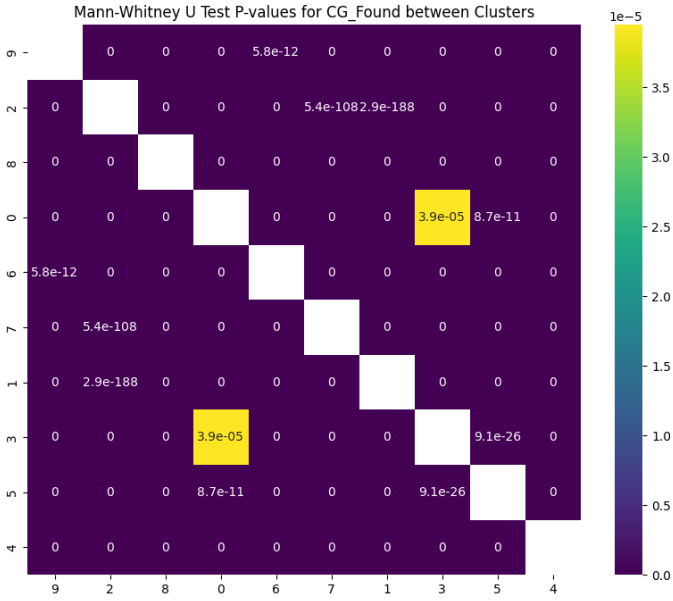

Fig. 20. Heat Map of 'CG'

Furthermore, the high-intensity clusters indicate areas with high motif frequency, possibly associated with conserved biological processes, regulatory function, or evolutionary constraint. Low-intensity clusters show places where there are fewer motif occurrences, which usually means that the motif is less important for function or that the region is variable. Lastly, comparative insights group differences that point to biological diversity, different gene expression, or differences in the structure of the bacterial genomes.

TABLE VII 'ATGC' Tables

(a) Mann-Whitney U tests and Bonferroni correction for 'ATGC' (Self Made)

Cluster_A	Cluster_B	U_statistic	P_value	P_value_corrected
9	2	5170620.5	0.000000e+00	True
9	8	0.0	0.000000e+00	True
9	0	19588.5	0.000000e+00	True
9	6	2236214.0	3.284059e-178	True
9	7	6485416.5	0.000000e+00	True
9	1	7338284.5	0.000000e+00	True
9	3	100726.5	0.000000e+00	True
9	5	3031408.0	2.991959e-58	True
9	4	78.5	0.000000e+00	True
2	8	0.0	0.000000e+00	True
2	0	0.5	0.000000e+00	True
2	6	356060.5	0.000000e+00	True
2	7	4213836.0	6.719851e-201	True
2	1	4485463.0	1.375281e-119	True
2	3	300.0	0.000000e+00	True
2	5	890342.5	0.000000e+00	True
2	4	0.0	0.000000e+00	True
8	0	5630357.0	0.000000e+00	True
8	6	7151912.0	0.000000e+00	True
8	7	6192392.0	0.000000e+00	True
8	1	7201376.0	0.000000e+00	True
8	3	7458625.0	0.000000e+00	True
8	5	7238297.5	0.000000e+00	True
8	4	2383153.5	2.031563e-21	True
0	6	6535125.0	0.000000e+00	True
0	7	5886904.0	0.000000e+00	True
0	1	6845639.5	0.000000e+00	True
0	3	5233839.5	2.865827e-195	True
0	5	6564566.5	0.000000e+00	True
0	4	404695.5	0.000000e+00	True
6	7	7153463.0	0.000000e+00	True
6	1	8230765.5	0.000000e+00	True
6	3	922137.5	0.000000e+00	True
6	5	4787961.5	4.179833e-18	True
6	4	1132.5	0.000000e+00	True
7	1	3157514.0	1.330352e-18	True
7	3	2.0	0.000000e+00	True
7	5	363729.0	0.000000e+00	True
7	4	0.0	0.000000e+00	True
1	3	1815.5	0.000000e+00	True
1	5	603597.5	0.000000e+00	True
1	4	0.0	0.000000e+00	True
3	5	7873925.5	0.000000e+00	True
3	4	145298.5	0.000000e+00	True
5	4	9419.5	0.000000e+00	True

(b) Mann-Whitney U Statistic for 'ATGC' (Self Made)

Comparison	U Statistic	P-value
Cluster 9 and Cluster 2	5170620.5	0.0
Cluster 9 and Cluster 8	0.0	0.0
Cluster 9 and Cluster 0	19588.5	0.0
Cluster 9 and Cluster 6	2236214.0	3.284058535537945e-178
Cluster 9 and Cluster 7	6485416.5	0.0
Cluster 9 and Cluster 1	7338284.5	0.0
Cluster 9 and Cluster 3	100726.5	0.0
Cluster 9 and Cluster 5	3031408.0	2.9919587909334675e-58
Cluster 9 and Cluster 4	78.5	0.0
Cluster 2 and Cluster 8	0.0	0.0
Cluster 2 and Cluster 0	0.5	0.0
Cluster 2 and Cluster 6	356060.5	0.0
Cluster 2 and Cluster 7	4213836.0	6.719850867031014e-201
Cluster 2 and Cluster 1	4485463.0	1.3752811472252888e-119
Cluster 2 and Cluster 3	300.0	0.0
Cluster 2 and Cluster 5	890342.5	0.0
Cluster 2 and Cluster 4	0.0	0.0
Cluster 8 and Cluster 0	5630357.0	0.0
Cluster 8 and Cluster 6	7151912.0	0.0
Cluster 8 and Cluster 7	6192392.0	0.0
Cluster 8 and Cluster 1	7201376.0	0.0
Cluster 8 and Cluster 3	7458625.0	0.0
Cluster 8 and Cluster 5	7238297.5	0.0
Cluster 8 and Cluster 4	2383153.5	2.0315632384410778e-21
Cluster 0 and Cluster 6	6535125.0	0.0
Cluster 0 and Cluster 7	5886904.0	0.0
Cluster 0 and Cluster 1	6845639.5	0.0
Cluster 0 and Cluster 3	5233839.5	2.865827170445793e-195
Cluster 0 and Cluster 5	6564566.5	0.0
Cluster 0 and Cluster 4	404695.5	0.0
Cluster 6 and Cluster 7	7153463.0	0.0
Cluster 6 and Cluster 1	8230765.5	0.0
Cluster 6 and Cluster 3	922137.5	0.0
Cluster 6 and Cluster 5	4787961.5	4.179832648062237e-18
Cluster 6 and Cluster 4	1132.5	0.0
Cluster 7 and Cluster 1	3157514.0	1.3303524571484816e-18
Cluster 7 and Cluster 3	2.0	0.0
Cluster 7 and Cluster 5	363729.0	0.0
Cluster 7 and Cluster 4	0.0	0.0
Cluster 1 and Cluster 3	1815.5	0.0
Cluster 1 and Cluster 5	603597.5	0.0
Cluster 1 and Cluster 4	0.0	0.0
Cluster 3 and Cluster 5	7873925.5	0.0
Cluster 3 and Cluster 4	145298.5	0.0
Cluster 5 and Cluster 4	9419.5	0.0

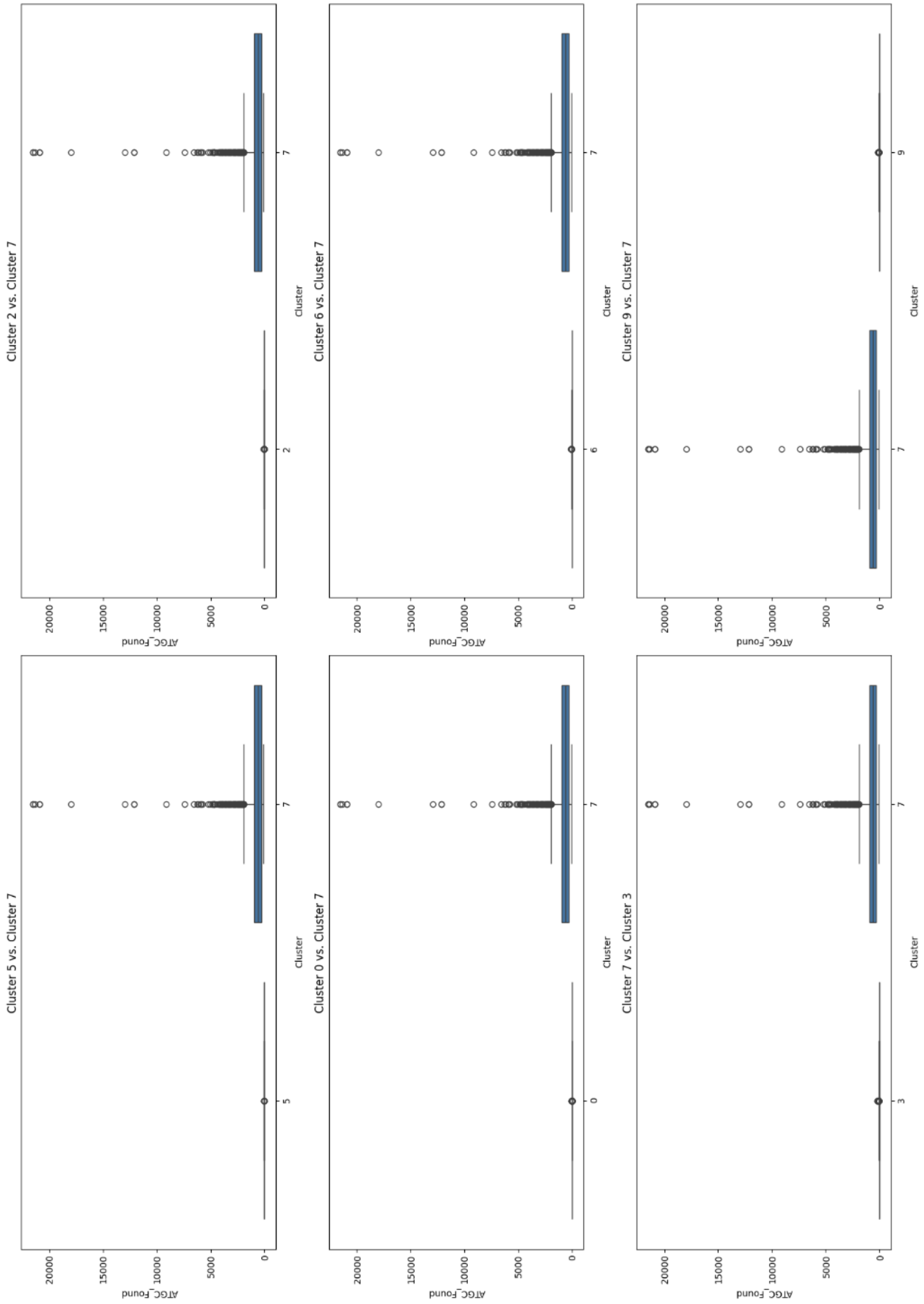

Fig. 21. Pairwise Cluster Plot of 'ATGC'

TABLE VIII 'TATAAA' Tables

(a) Mann-Whitney U tests and Bonferroni correction for 'TATAAA' (Self Made)

Cluster_A	Cluster_B	U_statistic	P_value	P_value_corrected
9	2	2333550.5	1.020660e-59	True
9	8	1855.0	0.000000e+00	True
9	0	3412.5	0.000000e+00	True
9	6	484948.0	0.000000e+00	True
9	7	4116330.5	9.027529e-54	True
9	1	5425715.5	5.995187e-249	True
9	3	704479.5	0.000000e+00	True
9	5	5368650.5	1.522856e-211	True
9	4	1268477.0	0.000000e+00	True
2	8	2329.5	0.000000e+00	True
2	0	4555.0	0.000000e+00	True
2	6	627126.0	0.000000e+00	True
2	7	4057820.0	2.163768e-201	True
2	1	5177320.0	0.000000e+00	True
2	3	834617.5	0.000000e+00	True
2	5	5142462.5	0.000000e+00	True
2	4	1333714.0	9.488347e-182	True
8	0	3745854.0	1.937339e-67	True
8	6	6718607.0	0.000000e+00	True
8	7	6192314.0	0.000000e+00	True
8	1	7201375.5	0.000000e+00	True
8	3	7169668.5	0.000000e+00	True
8	5	7238456.5	0.000000e+00	True
8	4	4666070.0	0.000000e+00	True
0	6	6370500.5	0.000000e+00	True
0	7	5886224.5	0.000000e+00	True
0	1	6846047.5	0.000000e+00	True
0	3	6819025.0	0.000000e+00	True
0	5	6881211.5	0.000000e+00	True
0	4	4004840.5	1.754035e-181	True
6	7	7022256.0	0.000000e+00	True
6	1	8274921.0	0.000000e+00	True
6	3	4635919.5	4.822753e-05	True
6	5	8302619.0	0.000000e+00	True
6	4	3359474.5	3.968423e-01	False
7	1	4283205.0	1.980328e-99	True
7	3	394700.5	0.000000e+00	True
7	5	4220871.5	1.119273e-68	True
7	4	886887.5	0.000000e+00	True
1	3	318059.5	0.000000e+00	True
1	5	4164129.5	9.943216e-07	True
1	4	808103.5	0.000000e+00	True
3	5	8506492.5	0.000000e+00	True
3	4	3368451.5	8.627895e-02	False
5	4	842149.0	0.000000e+00	True

(b) Mann-Whitney U Statistic for 'TATAAA' (Self Made)

Comparison	U Statistic	P-value
Cluster 9 and Cluster 2	2333550.5	1.020660221061724e-59
Cluster 9 and Cluster 8	1855.0	0.0
Cluster 9 and Cluster 0	3412.5	0.0
Cluster 9 and Cluster 6	484948.0	0.0
Cluster 9 and Cluster 7	4116330.5	9.027528873759478e-54
Cluster 9 and Cluster 1	5425715.5	5.995187104830138e-249
Cluster 9 and Cluster 3	704479.5	0.0
Cluster 9 and Cluster 5	5368650.5	1.522856278424965e-211
Cluster 9 and Cluster 4	1268477.0	0.0
Cluster 2 and Cluster 8	2329.5	0.0
Cluster 2 and Cluster 0	4555.0	0.0
Cluster 2 and Cluster 6	627126.0	0.0
Cluster 2 and Cluster 7	4057820.0	2.163768051791357e-201
Cluster 2 and Cluster 1	5177320.0	0.0
Cluster 2 and Cluster 3	834617.5	0.0
Cluster 2 and Cluster 5	5142462.5	0.0
Cluster 2 and Cluster 4	1333714.0	9.488346613992423e-182
Cluster 8 and Cluster 0	3745854.0	1.937339405575136e-67
Cluster 8 and Cluster 6	6718607.0	0.0
Cluster 8 and Cluster 7	6192314.0	0.0
Cluster 8 and Cluster 1	7201375.5	0.0
Cluster 8 and Cluster 3	7169668.5	0.0
Cluster 8 and Cluster 5	7238456.5	0.0
Cluster 8 and Cluster 4	4666070.0	0.0
Cluster 0 and Cluster 6	6370500.5	0.0
Cluster 0 and Cluster 7	5886224.5	0.0
Cluster 0 and Cluster 1	6846047.5	0.0
Cluster 0 and Cluster 3	6819025.0	0.0
Cluster 0 and Cluster 5	6881211.5	0.0
Cluster 0 and Cluster 4	4004840.5	1.7540348581057428e-181
Cluster 6 and Cluster 7	7022256.0	0.0
Cluster 6 and Cluster 1	8274921.0	0.0
Cluster 6 and Cluster 3	4635919.5	4.8227526459258916e-05
Cluster 6 and Cluster 5	8302619.0	0.0
Cluster 6 and Cluster 4	3359474.5	0.39684234025591325
Cluster 7 and Cluster 1	4283205.0	1.9803277257941056e-99
Cluster 7 and Cluster 3	394700.5	0.0
Cluster 7 and Cluster 5	4220871.5	1.1192726869205934e-68
Cluster 7 and Cluster 4	886887.5	0.0
Cluster 1 and Cluster 3	318059.5	0.0
Cluster 1 and Cluster 5	4164129.5	9.943215504578232e-07
Cluster 1 and Cluster 4	808103.5	0.0
Cluster 3 and Cluster 5	8506492.5	0.0
Cluster 3 and Cluster 4	3368451.5	0.08627894622339104
Cluster 5 and Cluster 4	842149.0	0.0

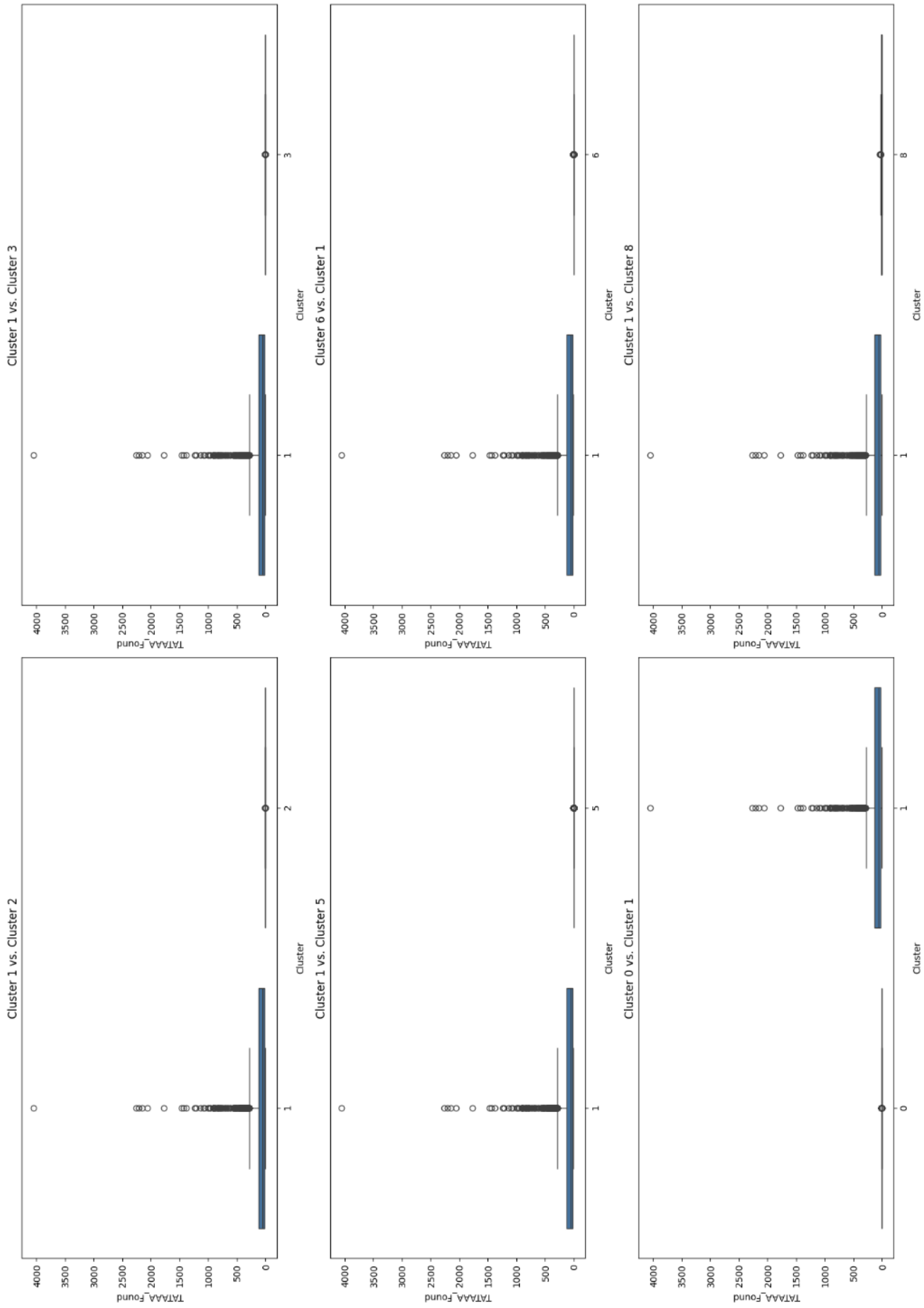

Fig. 22. Pairwise Cluster Plot of 'TATAAA'

TABLE IX CACACACACA Tables

(a) Mann-Whitney U tests and Bonferroni correction for 'CACACACACA' (Self Made)

Cluster_A	Cluster_B	U_statistic	P_value	P_value_corrected
9	2	3085752.5	5.542766e-02	False
9	8	3271310.0	1.090065e-24	True
9	0	3205334.0	2.338951e-06	True
9	6	3972305.0	1.679942e-01	False
9	7	3438999.0	1.384317e-01	False
9	1	4002544.0	1.000000e+00	False
9	3	4137245.0	9.430661e-04	True
9	5	4014914.5	1.756115e-02	False
9	4	3099497.5	2.301600e-11	True
2	8	2678784.5	1.132200e-18	True
2	0	2624693.0	9.172326e-04	True
2	6	3252698.0	4.633987e-01	False
2	7	2816001.0	5.704288e-01	False
2	1	3277456.0	4.866886e-02	False
2	3	3387773.0	7.565536e-02	False
2	5	3287594.5	5.405324e-01	False
2	4	2538056.5	9.400729e-08	True
8	0	2993078.5	8.844633e-12	True
8	6	3707993.0	1.606549e-24	True
8	7	3210081.5	2.532075e-21	True
8	1	3736096.0	4.824635e-26	True
8	3	3862568.0	1.166358e-19	True
8	5	3748012.0	2.753425e-22	True
8	4	2895394.5	4.621775e-06	True
0	6	3424670.0	2.518551e-05	True
0	7	2964870.5	1.131211e-04	True
0	1	3450720.0	1.177446e-06	True
0	3	3566970.0	4.835595e-02	False
0	5	3461445.0	1.513075e-03	False
0	4	2672535.5	1.192059e-02	False
6	7	3620394.0	8.848837e-01	False
6	1	4213664.0	1.559066e-01	False
6	3	4355482.0	9.493473e-03	False
6	5	4226693.0	1.621292e-01	False
6	4	3263021.0	1.619875e-10	True
7	1	3648736.0	1.272686e-01	False
7	3	3771554.0	1.967656e-02	False
7	5	3660025.0	2.314510e-01	False
7	4	2825588.5	3.531315e-09	True
1	3	4382560.0	6.652103e-04	True
1	5	4252976.0	1.452091e-02	False
1	4	3283280.0	5.982486e-12	True
3	5	4431193.0	1.775029e-01	False
3	4	3421051.0	4.269009e-06	True
5	4	3307065.5	2.277340e-08	True

(b) Mann-Whitney U Statistic for 'CACACACACA' (Self Made)

Comparison	U Statistic	P-value
Cluster 0 and Cluster 1	3450720.0	1.1774462239376818e-06
Cluster 0 and Cluster 2	2660355.0	0.0009172326424974167
Cluster 0 and Cluster 3	3566970.0	0.04835595005888699
Cluster 0 and Cluster 4	2672535.5	0.01192059192723852
Cluster 0 and Cluster 5	3461445.0	0.0015130751881095648
Cluster 0 and Cluster 6	3424670.0	2.5185506397009544e-05
Cluster 0 and Cluster 7	2964870.5	0.00011312111917424375
Cluster 0 and Cluster 8	2820944.5	8.844632760361638e-12
Cluster 0 and Cluster 9	3257565.0	2.338951385594754e-06
Cluster 1 and Cluster 2	3268720.0	0.04866886057447547
Cluster 1 and Cluster 3	4382560.0	0.0006652103251385104
Cluster 1 and Cluster 4	3283280.0	5.982485740991957e-12
Cluster 1 and Cluster 5	4252976.0	0.014520907496482516
Cluster 1 and Cluster 6	4207840.0	0.15590664603164595
Cluster 1 and Cluster 7	3642912.0	0.12726855827248848
Cluster 1 and Cluster 8	3465280.0	4.824634748918482e-26
Cluster 1 and Cluster 9	4002544.0	1.0
Cluster 2 and Cluster 3	3387773.0	0.07565535787371426
Cluster 2 and Cluster 4	2538056.5	9.400729438795074e-08
Cluster 2 and Cluster 5	3287594.5	0.5405324034823966
Cluster 2 and Cluster 6	3252698.0	0.4633986706642611
Cluster 2 and Cluster 7	2816001.0	0.5704288044203933
Cluster 2 and Cluster 8	2678784.5	1.1321997226327302e-18
Cluster 2 and Cluster 9	3093999.5	0.05542766253415305
Cluster 3 and Cluster 4	3421051.0	4.269008827314057e-06
Cluster 3 and Cluster 5	4431193.0	0.17750294633840558
Cluster 3 and Cluster 6	4384142.0	0.009493473224480214
Cluster 3 and Cluster 7	3795534.0	0.019676556343680946
Cluster 3 and Cluster 8	3610838.0	1.166358450200408e-19
Cluster 3 and Cluster 9	4170233.0	0.0009430660871921847
Cluster 4 and Cluster 5	3401618.5	2.2773399657943575e-08
Cluster 4 and Cluster 6	3365443.0	1.6198748456887917e-10
Cluster 4 and Cluster 7	2913579.5	3.531314991697631e-09
Cluster 4 and Cluster 8	2772721.5	4.621774523002538e-06
Cluster 4 and Cluster 9	3201210.5	2.30160018385287e-11
Cluster 5 and Cluster 6	4238191.0	0.16212922827257692
Cluster 5 and Cluster 7	3669183.0	0.23145098151772336
Cluster 5 and Cluster 8	3490459.0	2.753424857915655e-22
Cluster 5 and Cluster 9	4031408.5	0.017561149279769424
Cluster 6 and Cluster 7	3620394.0	0.88488369640284
Cluster 6 and Cluster 8	3443923.0	1.6065490248606712e-24
Cluster 6 and Cluster 9	3977803.0	0.16799416896371666
Cluster 7 and Cluster 8	2982310.5	2.5320746759978937e-21
Cluster 7 and Cluster 9	3444497.0	0.13843170078123315
Cluster 8 and Cluster 9	3526967.0	1.0900647228143703e-24

Fig. 23. Pairwise Cluster Plot of 'CACACACACA'

TABLE X 'AGGAGG' Tables

(a) Mann-Whitney U tests and Bonferroni correction for 'AGGAGG' (Self Made)

Cluster_A	Cluster_B	U_statistic	P_value	P_value_corrected
9	2	3743740.5	1.474308e-72	True
9	8	117375.0	0.000000e+00	True
9	0	812761.5	0.000000e+00	True
9	6	3493602.5	1.928343e-20	True
9	7	4214801.5	5.394204e-90	True
9	1	5000102.5	1.425457e-129	True
9	3	1782608.5	0.000000e+00	True
9	5	1999600.0	4.489732e-272	True
9	4	17316.0	0.000000e+00	True
2	8	46804.5	0.000000e+00	True
2	0	431659.0	0.000000e+00	True
2	6	2215886.5	4.389265e-139	True
2	7	2852862.5	1.085475e-01	False
2	1	3401712.5	2.228093e-07	True
2	3	980812.5	0.000000e+00	True
2	5	1226653.5	0.000000e+00	True
2	4	4186.0	0.000000e+00	True
8	0	4661350.5	1.198599e-289	True
8	6	6952161.5	0.000000e+00	True
8	7	6136294.0	0.000000e+00	True
8	1	7147714.5	0.000000e+00	True
8	3	6785689.5	0.000000e+00	True
8	5	6060693.5	0.000000e+00	True
8	4	1723743.0	2.995754e-121	True
0	6	5684581.5	0.000000e+00	True
0	7	5412566.5	0.000000e+00	True
0	1	6330832.5	0.000000e+00	True
0	3	4715534.5	7.194136e-97	True
0	5	4138972.5	1.119042e-37	True
0	4	376938.5	0.000000e+00	True
6	7	4815007.5	3.875493e-164	True
6	1	5694779.0	3.292588e-217	True
6	3	2362075.0	2.859444e-225	True
6	5	2469993.5	7.966639e-186	True
6	4	35494.0	0.000000e+00	True
7	1	3739465.0	3.205571e-04	True
7	3	1070049.0	0.000000e+00	True
7	5	1345550.5	0.000000e+00	True
7	4	6524.0	0.000000e+00	True
1	3	1176076.0	0.000000e+00	True
1	5	1502187.5	0.000000e+00	True
1	4	4368.5	0.000000e+00	True
3	5	4165888.0	7.813079e-05	True
3	4	168715.5	0.000000e+00	True
5	4	422588.5	0.000000e+00	True

(b) Mann-Whitney U Statistic for 'AGGAGG' (Self Made)

Comparison	U Statistic	P-value
Cluster 0 and Cluster 1	6330832.5	0.0
Cluster 0 and Cluster 2	4853389.0	0.0
Cluster 0 and Cluster 3	4715534.5	7.194136162840307e-97
Cluster 0 and Cluster 4	376938.5	0.0
Cluster 0 and Cluster 5	4138972.5	1.1190418405306124e-37
Cluster 0 and Cluster 6	5684581.5	0.0
Cluster 0 and Cluster 7	5412566.5	0.0
Cluster 0 and Cluster 8	1152672.5	1.1985994995453325e-289
Cluster 0 and Cluster 9	5650137.5	0.0
Cluster 1 and Cluster 2	3144463.5	2.2280931685484288e-07
Cluster 1 and Cluster 3	1176076.0	0.0
Cluster 1 and Cluster 4	4368.5	0.0
Cluster 1 and Cluster 5	1502187.5	0.0
Cluster 1 and Cluster 6	2726725.0	3.292588269134284e-217
Cluster 1 and Cluster 7	3552183.0	0.0003205571296261137
Cluster 1 and Cluster 8	53661.5	0.0
Cluster 1 and Cluster 9	3004985.5	1.4254565352795113e-129
Cluster 2 and Cluster 3	980812.5	0.0
Cluster 2 and Cluster 4	4186.0	0.0
Cluster 2 and Cluster 5	1226653.5	0.0
Cluster 2 and Cluster 6	2215886.5	4.389265495931372e-139
Cluster 2 and Cluster 7	2852862.5	0.10854752103236326
Cluster 2 and Cluster 8	46804.5	0.0
Cluster 2 and Cluster 9	2436011.5	1.4743083021145065e-72
Cluster 3 and Cluster 4	168715.5	0.0
Cluster 3 and Cluster 5	4165888.0	7.813079132628286e-05
Cluster 3 and Cluster 6	6377549.0	2.8594440014565136e-225
Cluster 3 and Cluster 7	6497039.0	0.0
Cluster 3 and Cluster 8	687716.5	0.0
Cluster 3 and Cluster 9	6524869.5	0.0
Cluster 4 and Cluster 5	6286095.5	0.0
Cluster 4 and Cluster 6	6592970.0	0.0
Cluster 4 and Cluster 7	5732644.0	0.0
Cluster 4 and Cluster 8	3944373.0	2.99575401819296e-121
Cluster 4 and Cluster 9	6283392.0	0.0
Cluster 5 and Cluster 6	5994890.5	7.966638790198116e-186
Cluster 5 and Cluster 7	5983657.5	0.0
Cluster 5 and Cluster 8	1177777.5	0.0
Cluster 5 and Cluster 9	6046723.0	4.489732146710449e-272
Cluster 6 and Cluster 7	4815007.5	3.875492668764464e-164
Cluster 6 and Cluster 8	199754.5	0.0
Cluster 6 and Cluster 9	4456505.5	1.9283425110329658e-20
Cluster 7 and Cluster 8	56098.0	0.0
Cluster 7 and Cluster 9	2668694.5	5.394203977960902e-90
Cluster 8 and Cluster 9	6680902.0	0.0

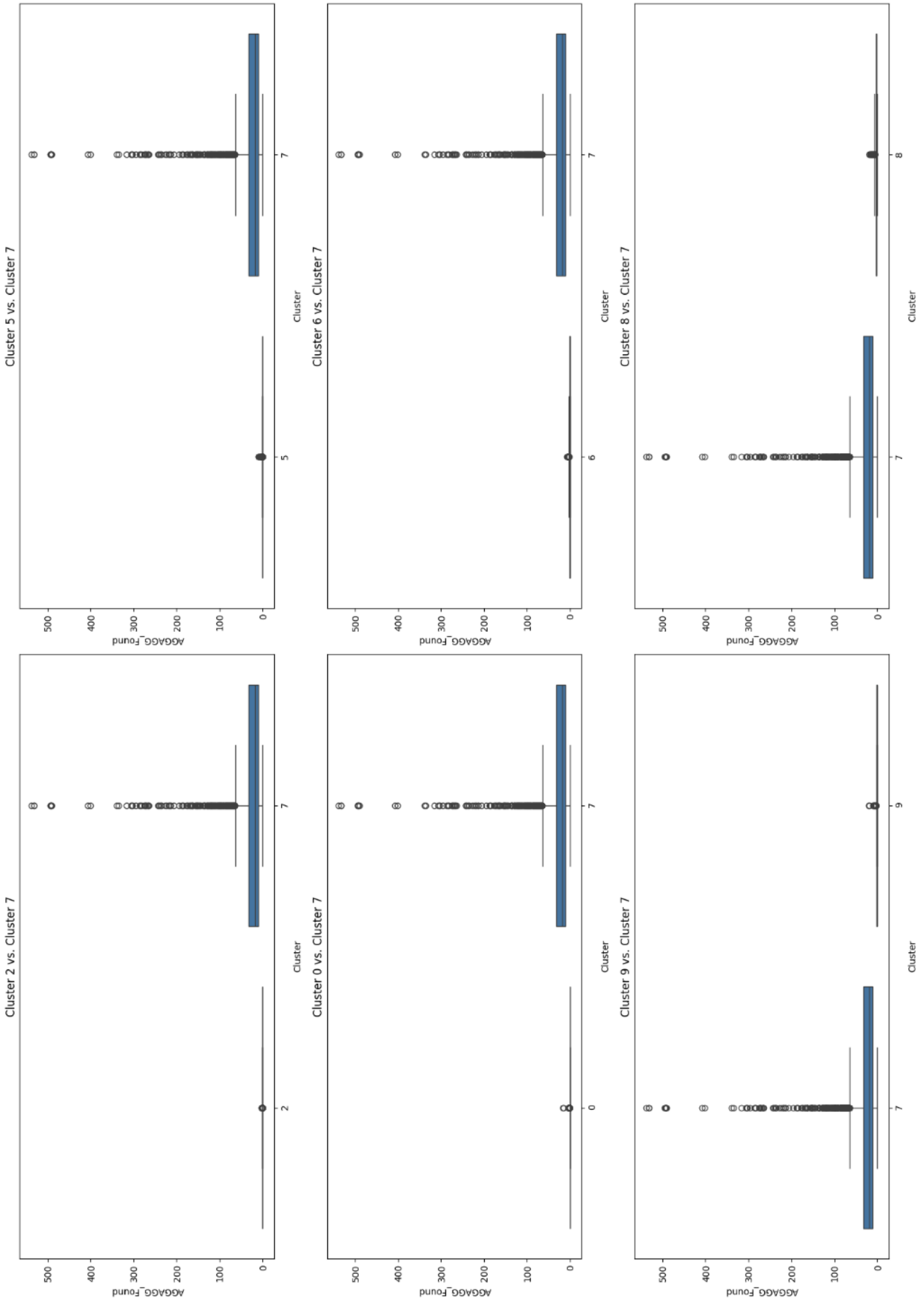

Fig. 24. Pairwise Cluster Plot of 'AGGAGG'

TABLE XI 'GAATTC' Tables

(a) Mann-Whitney U tests and Bonferroni correction for 'GAATTC' (Self Made)

Cluster_A	Cluster_B	U_statistic	P_value	P_value_corrected
9	2	3591163.5	2.708633e-38	True
9	8	5347.0	0.000000e+00	True
9	0	172705.5	0.000000e+00	True
9	6	2279976.0	5.543117e-197	True
9	7	4296707.5	2.144617e-106	True
9	1	5121070.0	7.223132e-160	True
9	3	1079339.5	0.000000e+00	True
9	5	3802901.0	2.034596e-05	True
9	4	55523.0	0.000000e+00	True
2	8	1314.0	0.000000e+00	True
2	0	82564.0	0.000000e+00	True
2	6	1461751.0	6.656768e-306	True
2	7	3066627.0	1.795245e-20	True
2	1	3673605.0	4.349071e-46	True
2	3	623908.0	0.000000e+00	True
2	5	2628716.5	9.947186e-57	True
2	4	31982.5	0.000000e+00	True
8	0	4732527.5	0.000000e+00	True
8	6	7057297.5	0.000000e+00	True
8	7	6191752.0	0.000000e+00	True
8	1	7200948.0	0.000000e+00	True
8	3	7131678.0	0.000000e+00	True
8	5	7182688.5	0.000000e+00	True
8	4	2514552.5	1.633281e-11	True
0	6	6090310.5	0.000000e+00	True
0	7	5819340.0	0.000000e+00	True
0	1	6776878.0	0.000000e+00	True
0	3	5540881.5	2.073570e-274	True
0	5	6506065.0	0.000000e+00	True
0	4	752970.0	0.000000e+00	True
6	7	5830079.0	0.000000e+00	True
6	1	6868129.5	0.000000e+00	True
6	3	2667224.0	9.849296e-153	True
6	5	5671579.5	1.516546e-128	True
6	4	162235.0	0.000000e+00	True
7	1	3766489.0	1.094133e-06	True
7	3	570606.0	0.000000e+00	True
7	5	2642662.5	1.690425e-129	True
7	4	29267.5	0.000000e+00	True
1	3	614575.0	0.000000e+00	True
1	5	2953160.5	1.135968e-185	True
1	4	31501.0	0.000000e+00	True
3	5	7278679.0	0.000000e+00	True
3	4	342916.5	0.000000e+00	True
5	4	96691.5	0.000000e+00	True

(b) Mann-Whitney U Statistic for 'GAATTC' (Self Made)

Comparison	U Statistic	P-value
Cluster 0 and Cluster 1	6776878.0	0.0
Cluster 0 and Cluster 2	5202484.0	0.0
Cluster 0 and Cluster 3	5540881.5	2.0735702035072612e-274
Cluster 0 and Cluster 4	752970.0	0.0
Cluster 0 and Cluster 5	6506065.0	0.0
Cluster 0 and Cluster 6	6090310.5	0.0
Cluster 0 and Cluster 7	5819340.0	0.0
Cluster 0 and Cluster 8	1081495.5	0.0
Cluster 0 and Cluster 9	6290193.5	0.0
Cluster 1 and Cluster 2	2872571.0	4.349070873736735e-46
Cluster 1 and Cluster 3	614575.0	0.0
Cluster 1 and Cluster 4	31501.0	0.0
Cluster 1 and Cluster 5	2953160.5	1.1359675530744989e-185
Cluster 1 and Cluster 6	1553374.5	0.0
Cluster 1 and Cluster 7	3525159.0	1.0941334322852506e-06
Cluster 1 and Cluster 8	428.0	0.0
Cluster 1 and Cluster 9	2884018.0	7.22313244547663e-160
Cluster 2 and Cluster 3	623908.0	0.0
Cluster 2 and Cluster 4	31982.5	0.0
Cluster 2 and Cluster 5	2628716.5	9.947186077997436e-57
Cluster 2 and Cluster 6	1461751.0	6.656767900885095e-306
Cluster 2 and Cluster 7	3066627.0	1.7952446209218096e-20
Cluster 2 and Cluster 8	1314.0	0.0
Cluster 2 and Cluster 9	2588588.5	2.7086331494017764e-38
Cluster 3 and Cluster 4	342916.5	0.0
Cluster 3 and Cluster 5	7278679.0	0.0
Cluster 3 and Cluster 6	6072400.0	9.849295504745095e-153
Cluster 3 and Cluster 7	6996482.0	0.0
Cluster 3 and Cluster 8	341728.0	0.0
Cluster 3 and Cluster 9	7228138.5	0.0
Cluster 4 and Cluster 5	6611992.5	0.0
Cluster 4 and Cluster 6	6466229.0	0.0
Cluster 4 and Cluster 7	5709900.5	0.0
Cluster 4 and Cluster 8	3153563.5	1.633281243190368e-11
Cluster 4 and Cluster 9	6245185.0	0.0
Cluster 5 and Cluster 6	2793304.5	1.5165458706733963e-128
Cluster 5 and Cluster 7	4686545.5	1.6904248316043587e-129
Cluster 5 and Cluster 8	55782.5	0.0
Cluster 5 and Cluster 9	4243422.0	2.0345958850291696e-05
Cluster 6 and Cluster 7	5830079.0	0.0
Cluster 6 and Cluster 8	94618.5	0.0
Cluster 6 and Cluster 9	5670132.0	5.543116830742381e-197
Cluster 7 and Cluster 8	640.0	0.0
Cluster 7 and Cluster 9	2586788.5	2.1446169362489658e-106
Cluster 8 and Cluster 9	6792930.0	0.0

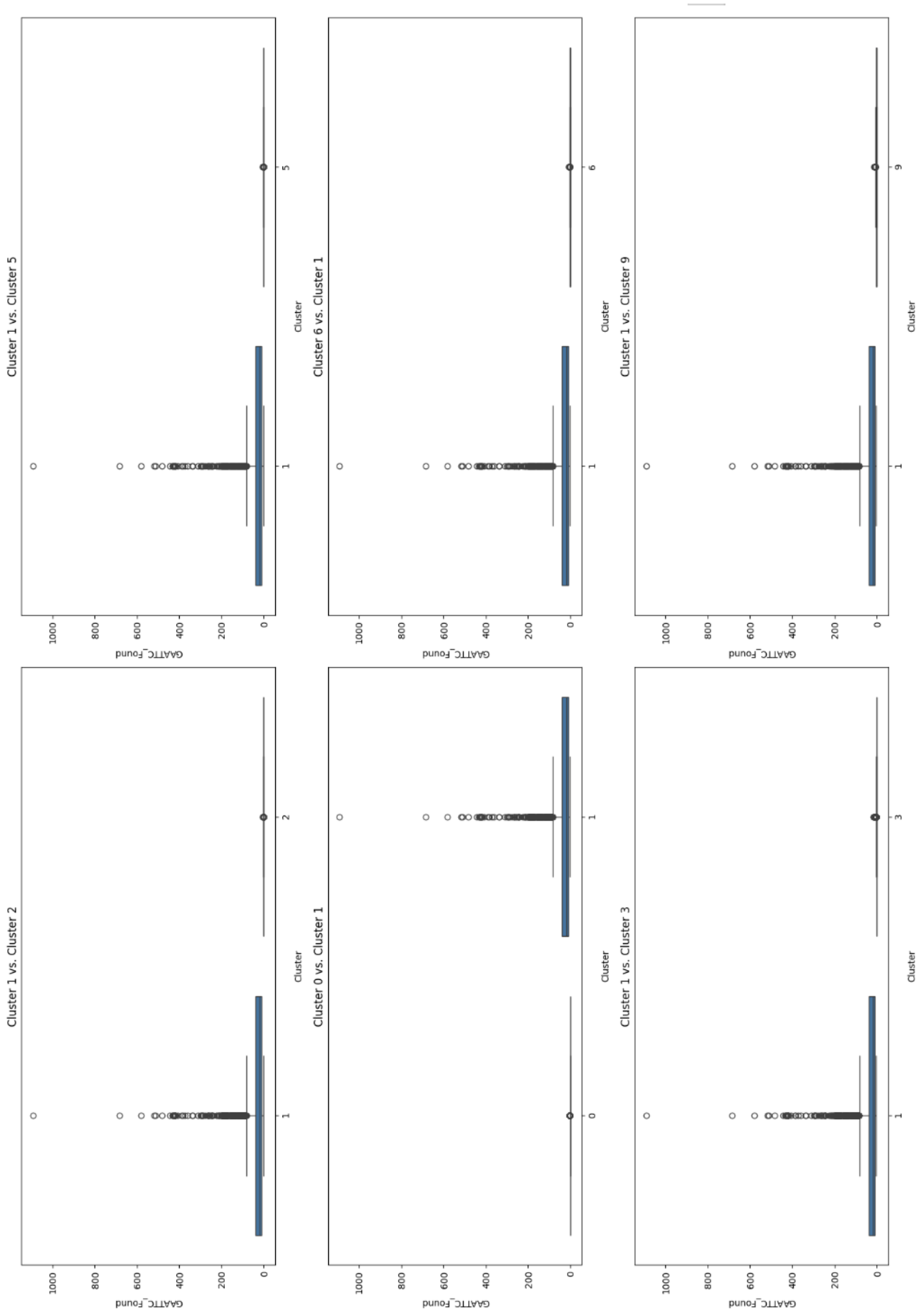

Fig. 25. Pairwise Cluster Plot of 'GAATTC'

TABLE XII 'CG'Tables

(a) Mann-Whitney U tests and Bonferroni correction for 'CG' (Self Made)

Cluster_A	Cluster_B	U_statistic	P_value	P_value_corrected
9	2	6027772.5	0.000000e+00	True
9	8	1209.0	0.000000e+00	True
9	0	274459.0	0.000000e+00	True
9	6	4395999.5	5.769400e-12	True
9	7	6852394.5	0.000000e+00	True
9	1	7405889.0	0.000000e+00	True
9	3	285737.0	0.000000e+00	True
9	5	898880.0	0.000000e+00	True
9	4	0.0	0.000000e+00	True
2	8	0.0	0.000000e+00	True
2	0	1592.0	0.000000e+00	True
2	6	381179.0	0.000000e+00	True
2	7	3856358.0	5.382252e-108	True
2	1	1720590.0	2.923567e-188	True
2	3	1.0	0.000000e+00	True
2	5	554.0	0.000000e+00	True
2	4	0.0	0.000000e+00	True
8	0	5704421.0	0.000000e+00	True
8	6	7150784.0	0.000000e+00	True
8	7	6192392.0	0.000000e+00	True
8	1	7201376.0	0.000000e+00	True
8	3	7208062.5	0.000000e+00	True
8	5	6860442.0	0.000000e+00	True
8	4	910721.0	0.000000e+00	True
0	6	6503477.0	0.000000e+00	True
0	7	5886665.5	0.000000e+00	True
0	1	6839282.5	0.000000e+00	True
0	3	3320501.0	3.948255e-05	True
0	5	3797622.5	8.738285e-11	True
0	4	661.0	0.000000e+00	True
6	7	7120146.0	0.000000e+00	True
6	1	7250323.0	0.000000e+00	True
6	3	334758.5	0.000000e+00	True
6	5	897207.5	0.000000e+00	True
6	4	0.0	0.000000e+00	True
7	1	765610.5	0.000000e+00	True
7	3	0.0	0.000000e+00	True
7	5	0.0	0.000000e+00	True
7	4	0.0	0.000000e+00	True
1	3	545.5	0.000000e+00	True
1	5	17717.5	0.000000e+00	True
1	4	0.0	0.000000e+00	True
3	5	5117748.0	9.090032e-26	True
3	4	7060.5	0.000000e+00	True
5	4	29304.5	0.000000e+00	True

(b) Mann-Whitney U Statistic for 'CG' (Self Made)

Comparison	U Statistic	P-value
Cluster 0 and Cluster 1	6839282.5	0.0
Cluster 0 and Cluster 2	5283456.0	0.0
Cluster 0 and Cluster 3	3320501.0	3.948254917325167e-05
Cluster 0 and Cluster 4	661.0	0.0
Cluster 0 and Cluster 5	3797622.5	8.738284827351792e-11
Cluster 0 and Cluster 6	6503477.0	0.0
Cluster 0 and Cluster 7	5886665.5	0.0
Cluster 0 and Cluster 8	109602.0	0.0
Cluster 0 and Cluster 9	6188440.0	0.0
Cluster 1 and Cluster 2	4825586.0	2.9235670967905904e-188
Cluster 1 and Cluster 3	545.5	0.0
Cluster 1 and Cluster 4	0.0	0.0
Cluster 1 and Cluster 5	17717.5	0.0
Cluster 1 and Cluster 6	1171181.0	0.0
Cluster 1 and Cluster 7	6526037.5	0.0
Cluster 1 and Cluster 8	0.0	0.0
Cluster 1 and Cluster 9	599199.0	0.0
Cluster 2 and Cluster 3	1.0	0.0
Cluster 2 and Cluster 4	0.0	0.0
Cluster 2 and Cluster 5	554.0	0.0
Cluster 2 and Cluster 6	381179.0	0.0
Cluster 2 and Cluster 7	3856358.0	5.3822516983372755e-108
Cluster 2 and Cluster 8	0.0	0.0
Cluster 2 and Cluster 9	151979.5	0.0
Cluster 3 and Cluster 4	7060.5	0.0
Cluster 3 and Cluster 5	5117748.0	9.090031744869921e-26
Cluster 3 and Cluster 6	8404865.5	0.0
Cluster 3 and Cluster 7	7567088.0	0.0
Cluster 3 and Cluster 8	265343.5	0.0
Cluster 3 and Cluster 9	8021741.0	0.0
Cluster 4 and Cluster 5	6679379.5	0.0
Cluster 4 and Cluster 6	6628464.0	0.0
Cluster 4 and Cluster 7	5739168.0	0.0
Cluster 4 and Cluster 8	4757395.0	0.0
Cluster 4 and Cluster 9	6300708.0	0.0
Cluster 5 and Cluster 6	7567676.5	0.0
Cluster 5 and Cluster 7	7329208.0	0.0
Cluster 5 and Cluster 8	378029.0	0.0
Cluster 5 and Cluster 9	7147443.0	0.0
Cluster 6 and Cluster 7	7120146.0	0.0
Cluster 6 and Cluster 8	1132.0	0.0
Cluster 6 and Cluster 9	3554108.5	5.769400480754909e-12
Cluster 7 and Cluster 8	0.0	0.0
Cluster 7 and Cluster 9	31101.5	0.0
Cluster 8 and Cluster 9	6797068.0	0.0

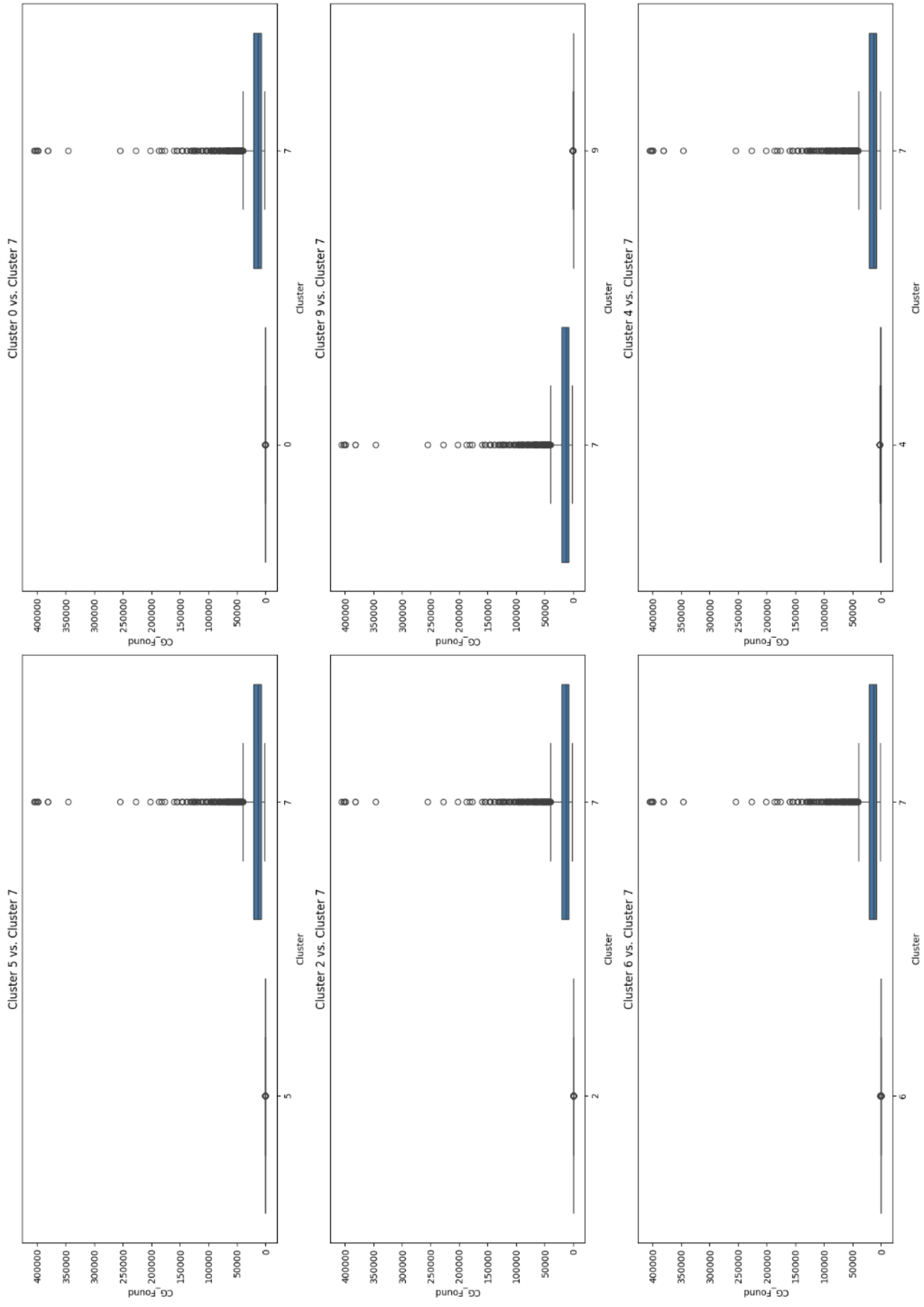

Fig. 26. Pairwise Cluster Plot of 'CG'

V. Discussion

Understanding bacterial genomes is easier when we look at the building blocks of DNA, especially in important areas like the TATA Box, CAAT Box, initiation codon, and the EcoRI site ('GAATTC') and 'AGGAGG.' These differences show how bacteria naturally adapt to their environment and fight other organisms. Specific gene promoter patterns, such as TATA and CAAT boxes, can alter gene usage, impacting the survival and energy utilization of bacteria. Furthermore, differences in the start of gene translation can change the types of proteins bacteria make, affecting their growth and ability to cause disease. Changes in the EcoRI site ('GAATTC') can help harmful bacteria avoid the body's defenses, which are essential for survival. Differences in CpG islands can lead to changes in gene control and expression. Bacteria need specific traits to survive in different places. We can learn how genetic differences lead to various bacteria by studying these traits. This research could help us find new ways to fight bacteria and help sick people.

VI. Conclusion

The investigation examines how bacteria's genetic material differences can help us understand their evolution and biology. It found essential variations in genetic patterns among different groups of bacteria. These differences can help us understand how bacteria develop resistance, cause disease, and adapt to various environments. This study improves our understanding of how bacteria evolve and could lead to better ways to treat infections and control harmful bacteria. It emphasizes the importance of using advanced computer analysis to study the genetic makeup of bacteria and how it relates to their characteristics.

References

[1] R. Yamada, D. Okada, J. Wang, T. Basak, S. Koyama,*Interpretation of omics data analyses*, J. Hum. Genet. 66 (2021) 93–102.

[2] K.R. Kumar, M.J. Cowley, R.L. Davis, *Next-generation sequencing and emerging technologies*, Semin. Thromb. Hemost. 45 (2019) 661–673.

[3] C.M. Kobras, A.K. Fenton, S.K. Sheppard, *Next-generation microbiology: from comparative genomics to gene function* , Genome Biol. 22 (2021).

[4] A. Almeida, S. Nayfach, M. Boland, F. Strozzi, M. Beracochea, Z.J. Shi, K.S. Pollard, E. Sakharova, D.H. Parks, P. Hugenholtz, N. Segata, N.C. Kyrpides, R.D. Finn, *A unified catalog of 204,938 reference genomes from the human gut microbiome* , Nat. Biotechnol. 39 (2021) 105–114.

[5] D.M.P. De Oliveira, B.M. Forde, T.J. Kidd, P.N.A. Harris, M.A. Schembri, S.A. Beatson, D.L. Paterson, M.J. Walker,*Antimicrobial resistance in ESKAPE pathogens, Clin. Microbiol* . Rev. 33 (2020).

[6] J. Mosquera-Rendón, C.X. Moreno-Herrera, J. Robledo, U. Hurtado-Páez, *Genome-wide association studies (GWAS) approaches for the detection of genetic variants associated with antibiotic resistance: A systematic review, Microorganisms.* 11 (2023) 2866.

[7] R.Y.K. Chang, S.C. Nang, H.-K. Chan, J. Li, *Novel antimicrobial agents for combating antibiotic-resistant bacteria, Adv. Drug Deliv* . Rev. 187 (2022) 114378.

[8] J.S. Johnson, D.J. Spakowicz, B.-Y. Hong, L.M. Petersen, P. Demkowicz, L. Chen, S.R. Leopold, B.M. Hanson, H.O. Agresta, M. Gerstein, E. Sodergren, G.M. Weinstock, *Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis*, Nat. Commun. 10 (2019).

[9] L. Stewart, A. Ford, V. Sangal, J. Jeukens, B. Boyle, I. Kukavica-Ibrulj, S. Caim, L. Crossman, P.A. Hoskisson, R. Levesque, N.P. Tucker, *Draft genomes of 12 host-adapted and environmental isolates of Pseudomonas aeruginosaand their positions in the core genome phylogeny*, Pathog. Dis. 71 (2014) 20–25.

[10] R.L. Marvig, L.M. Sommer, L. Jelsbak, S. Molin, H.K. Johansen, *Evolutionary insight from whole-genome sequencing of Pseudomonas aeruginosa from cystic fibrosis patients, Future Microbiol.* 10 (2015) 599–611.

[11] U. Muthukumarasamy, M. Preusse, A. Kordes, M. Koska, M. Schniederjans, A. Khaledi, S. Häussler, *Single-nucleotide polymorphism-based genetic diversity analysis of clinical Pseudomonas aeruginosa isolates* , Genome Biol. Evol. 12 (2020) 396–406.

[12] M. Orsini, G. Cuccuru, P. Uva, G. Fotia,*Bacterial genomic data analysis in the next-generation sequencing era*, in: Methods in Molecular Biology, Springer New York, New York, NY, 2016: pp. 407–422.

[13] J. Wagner, P. Coupland, H.P. Browne, T.D. Lawley, S.C. Francis, J. Parkhill, *Evaluation of PacBio sequencing for full-length bacterial 16S rRNA gene classification*, BMC Microbiol. 16 (2016).

[14] J.P. Earl, N.D. Adappa, J. Krol, A.S. Bhat, S. Balashov, R.L. Ehrlich, J.N. Palmer, A.D. Workman, M. Blasetti, B. Sen, J. Hammond, N.A. Cohen, G.D. Ehrlich, J.C. Mell, *Species-level bacterial community profiling of the healthy sinonasal microbiome using Pacific Biosciences sequencing of full-length 16S rRNA genes*, Microbiome. 6 (2018).

[15] T.W. Whon, N.-R. Shin, J.Y. Kim, S.W. Roh, *Omics in gut microbiome analysis*, J. Microbiol. 59 (2021) 292–297.

[16] E. Bolyen et al., *"Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2*," Nat. Biotechnol., vol. 37, no. 8, pp. 852–857, 2019.

[17] H. Dai, Y. Guan,*The Nubeam reference-free approach to analyze metagenomic sequencing reads*, Genome Res. 30 (2020) 1364–1375.

[18] X. Dai, L. Shen, *Advances and trends in omics technology development* , Front. Med. (Lausanne). 9 (2022).

[19] R. Patil, R. Satpute, D. Nalage,*The application of omics technologies to toxicology*, Toxicol. Adv. 5 (2023) 6.

[20] L. Cantini, P. Zakeri, C. Hernandez, A. Naldi, D. Thieffry, E. Remy, A. Baudot, *Benchmarking joint multi-omics dimensionality reduction approaches for cancer study*, bioRxiv. (2020).

[21] M. Krassowski, V. Das, S.K. Sahu, B.B. Misra, *State of the field in multi-omics research: From computational needs to data mining and sharing*, Front. Genet. 11 (2020).

[22] F.P. Breitwieser, D.N. Baker, S.L. Salzberg, KrakenUniq: *confident and fast metagenomics classification using unique k-mer counts* , Genome Biol. 19 (2018).

[23] A.R. Wattam, T. Brettin, J.J. Davis, S. Gerdes, R. Kenyon, D. Machi, C. Mao, R. Olson, R. Overbeek, G.D. Pusch, M.P. Shukla, R. Stevens, V. Vonstein, A. Warren, F. Xia, H. Yoo, *Assembly, annotation, and comparative genomics in PATRIC, the all bacterial bioinformatics resource center, in: Comparative Genomics*, Springer New York, New York, NY, 2018: pp. 79–101.