Supplement

Contents

1 Dataset selection process for Bacteria, Archaea, *E. coli* and MAGs

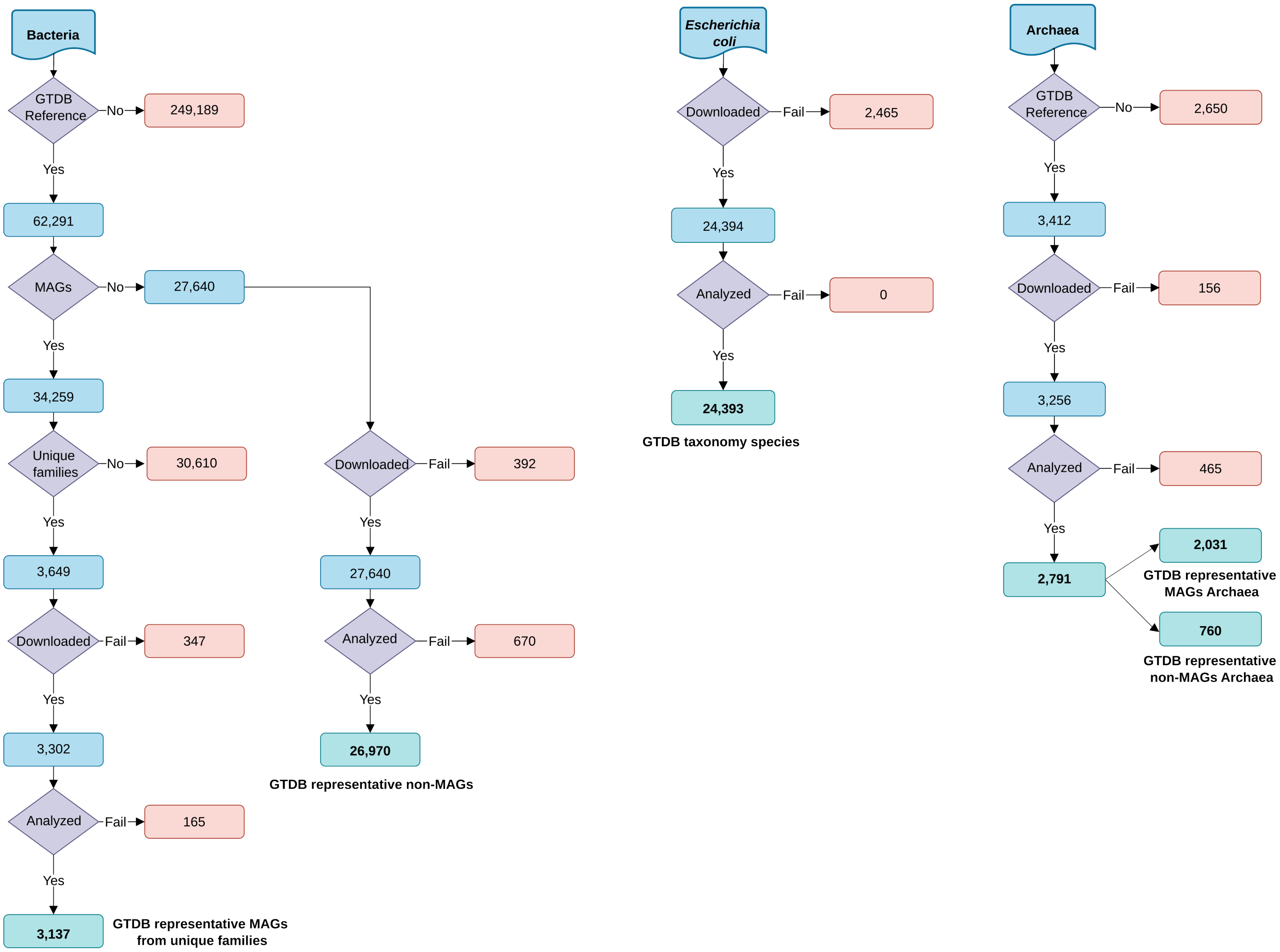


Figure 1: Flowchart of the sample requirement to the analysis. The blue boxes at the top represent the groups being investigated in the analysis. Purple rhombuses indicate decision points where samples may be excluded from the analysis. Green boxes show the number of samples that pass each decision point, while red boxes show the number of samples that are rejected or fail at each stage.

2 Dataset composition

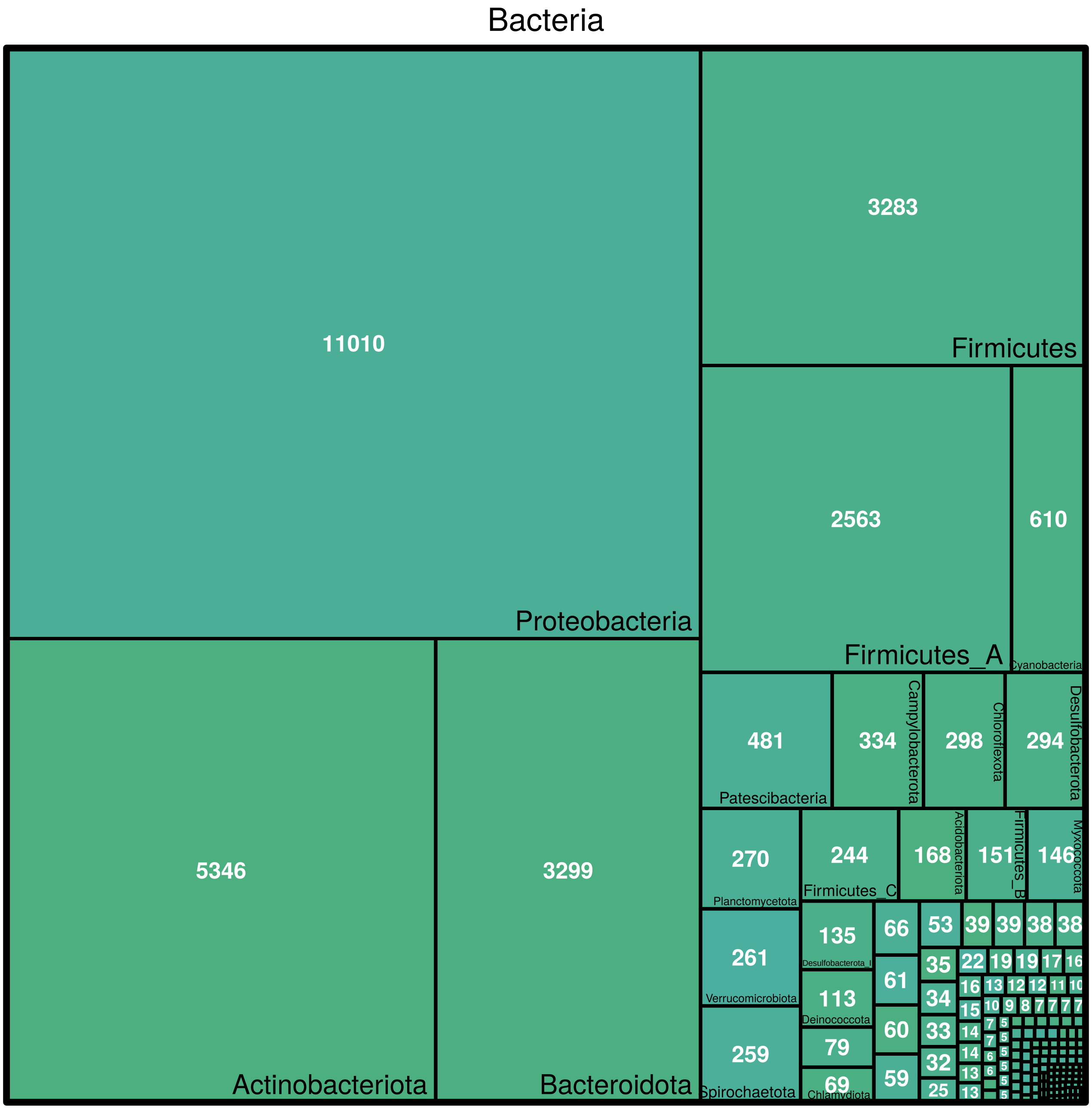


Figure 2: The Composition of the Bacteria kingdom group, divided by their phyla. *E. coli* baseline annotation group is excluded from visualization. Phyla nomenclature based on GTDB database version 107. Only pylas wit more than 100 representatives are labeled.

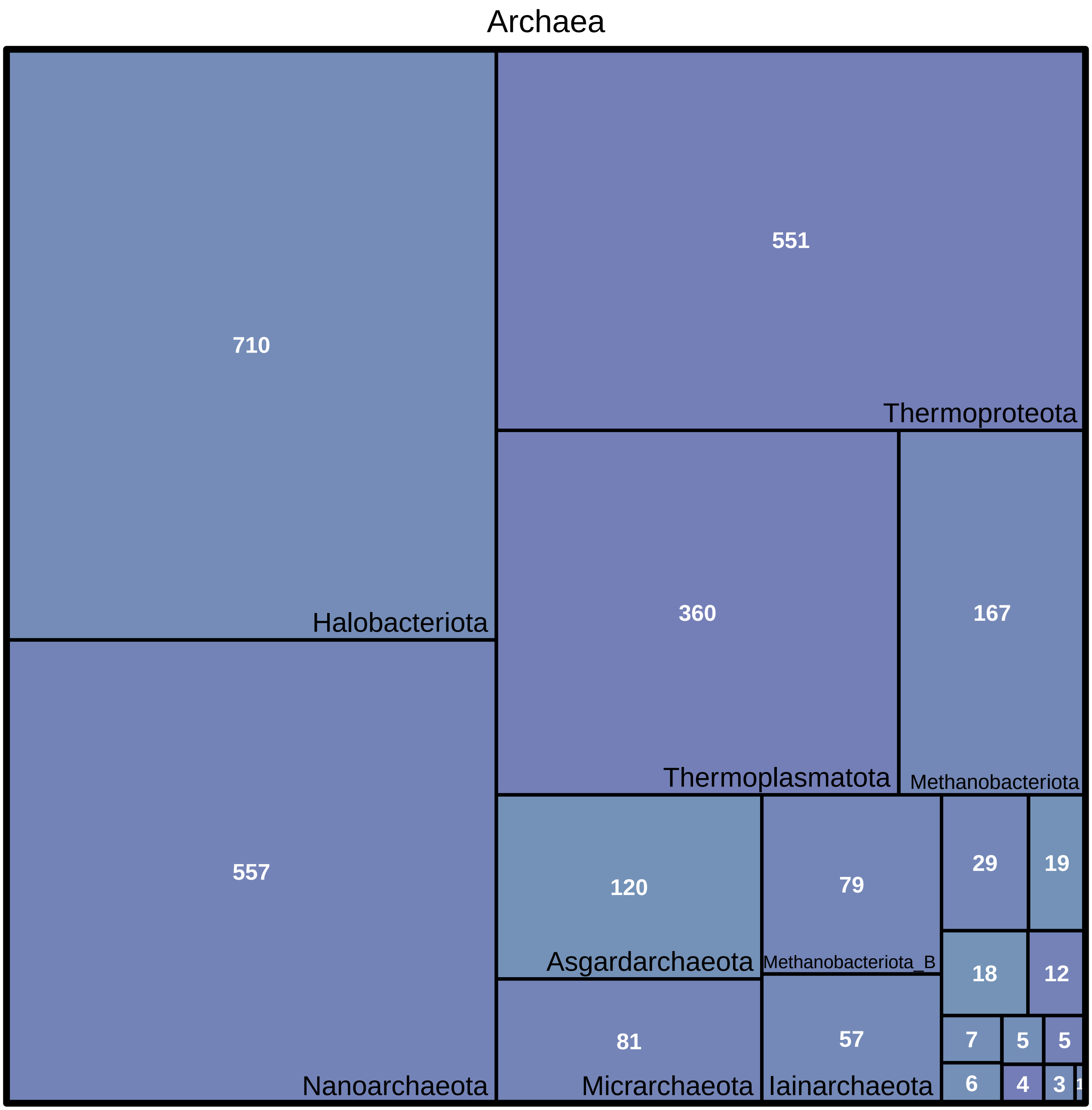


Figure 3: The Composition of the Archaea kingdom group, divided by their phyla. Phyla nomenclature based on GTDB database version 107. Only pylas wit more than 50 representatives are labeled.

3 Baseline annotation (*E. coli*)

3.1 Data description for *E. coli* strains

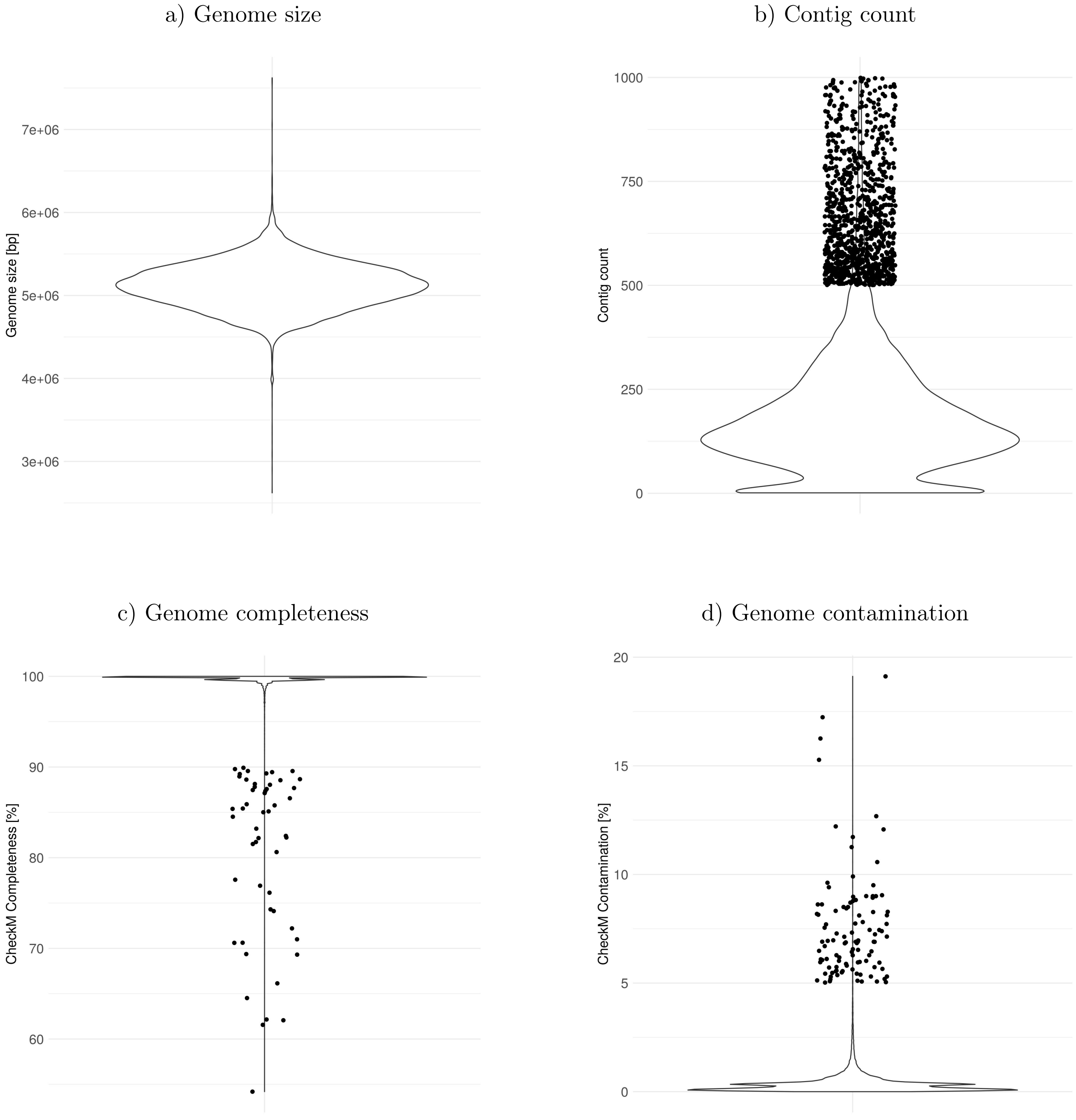


Figure 4: Basic metrics characterizing the *Escherichia coli* subgroup within the analyzed samples. Each panel presents a distinct attribute reported in GTDB dataset: (A) Genome size derived from genome_size variable, (B) Contig count derived from contig_count variable, (C) Genome completeness derived from GTDB checkm_completeness variable, and (D) Genome contamination derived from GTDB checkm_contamination variable.

3.2 Total feature count

Table 1: Comparative results of total feature count and one-way ANOVA Tukey HSD post hoc test. The differences on example of Prokka vs Bakta means that on average Prokka found 349.60890 less features than Bakta.

Comparison	Mean difference [N]	Lower end point of CI 95%	Upper end point of CI 95%	adj P-value
Prokka vs Bakta	-349.60890	-358.58465	-340.63316	< 0.001
Prokka vs EggNOG-mapper	-21.41684	-30.39259	-12.44109	< 0.001
Prokka vs PGAP	-561.75821	-570.73396	-552.78246	< 0.001
EggNOG-mapper vs Bakta	-328.19206	-337.16781	-319.21632	< 0.001
PGAP vs Bakta	212.14931	203.17356	221.12505	< 0.001
PGAP vs EggNOG-mapper	540.34137	531.36562	549.31712	< 0.001

3.3 Undescribed feature count

Table 2: Comparative results of undescribed feature count d and one-way ANOVA Tukey HSD post hoc test. The differences on example of Prokka vs Bakta means that on average Prokka had 1366.9815 more undescribed features than Bakta.

Comparison	Mean difference [N]	Lower end point of CI 95%	Upper end point of CI 95%	adj P-value
Prokka vs Bakta	1366.9815	1362.7445	1371.2185	< 0.001
Prokka vs EggNOG-mapper	971.2406	967.0036	975.4776	< 0.001
Prokka vs PGAP	663.2449	659.0079	667.4820	< 0.001
EggNOG-mapper vs Bakta	395.7409	391.5039	399.9779	< 0.001
PGAP vs Bakta	703.7365	699.4995	707.9735	< 0.001
PGAP vs EggNOG-mapper	307.9956	303.7586	312.2326	< 0.001

3.4 Total feature length

Table 3: Comparative results of total features length and one-way ANOVA Tukey HSD post hoc test.The differences on example of Prokka vs Bakta means that on average Prokka had 4.5229879 aa longer features than Bakta.

Comparison	Mean difference [aa]	Lower end point of CI 95%	Upper end point of CI 95%	adj P-value
Prokka vs Bakta	4.5229879	4.2980074	4.7479685	< 0.001
Prokka vs EggNOG-mapper	3.0849649	2.8599844	3.3099455	< 0.001
Prokka vs PGAP	3.9391125	3.7141320	4.1640931	< 0.001
EggNOG-mapper vs Bakta	1.4380230	1.2130425	1.6630036	< 0.001
PGAP vs Bakta	0.5838754	0.3588949	0.8088560	< 0.001
PGAP vs EggNOG-mapper	-0.8541476	-1.0791282	-0.6291671	< 0.001

3.5 Undescribed feature length

Table 4: Comparative results of undescribed features lenght and one-way ANOVA Tukey HSD post hoc test.The differences on example of Prokka vs Bakta means that on average Prokka had 128.774889 aa longer features than Bakta.

Comparison	Mean difference [aa]	Lower end point of CI 95%	Upper end point of CI 95%	adj P-value
Prokka vs Bakta	128.774889	127.7020696	129.8477080	< 0.001
EggNOG-mapper vs Bakta	128.622926	127.5501071	129.6957460	< 0.001
PGAP vs Bakta	14.235710	13.1628907	15.3085290	< 0.001
PGAP vs EggNOG-mapper	-114.387216	-115.4600358	-113.3143970	< 0.001
Prokka vs EggNOG-mapper	0.151963	-0.9208569	1.2247820	0.9835263
Prokka vs PGAP	114.539179	113.4663595	115.6119980	< 0.001

3.6 Total annotated coding space

Table 5: Comparative results of total annotated coding space and one-way ANOVA Tukey HSD post hoc test of proportion.The differences on example of Prokka vs Bakta means that on average Prokka had 39820.011 bp smaller total coding space.

Comparison	Mean difference [bp]	Lower end point of CI 95%	Upper end point of CI 95%	adj P-value
Prokka vs Bakta	-39820.011	-45611.436	-34028.585	< 0.001
Prokka vs EggNOG-mapper	-41402.180	-47193.606	-35610.754	< 0.001
Prokka vs PGAP	-251480.428	-257271.854	-245689.002	< 0.001
EggNOG-mapper vs Bakta	1582.169	-4209.256	7373.595	0.896
PGAP vs Bakta	211660.417	205868.992	217451.843	< 0.001
PGAP vs EggNOG-mapper	210078.248	204286.822	215869.674	< 0.001

3.7 Described annotated coding space

Table 6: Comparative results of described coding space and one-way ANOVA Tukey HSD post hoc test. The differences on example of Prokka vs Bakta means that on average Prokka had 870500.33 bp smaller described coding space.

Comparison	Mean difference [bp]	Lower end point of CI 95%	Upper end point of CI 95%	adj P-value
Prokka vs Bakta	-870500.33	-874927.12	-866073.54	< 0.001
Prokka vs EggNOG-mapper	-733250.20	-737676.99	-728823.41	< 0.001
Prokka vs PGAP	-816710.67	-821137.46	-812283.88	< 0.001
EggNOG-mapper vs Bakta	-137250.13	-141676.91	-132823.34	< 0.001
PGAP vs Bakta	-53789.66	-58216.44	-49362.87	< 0.001
PGAP vs EggNOG-mapper	83460.47	79033.68	87887.26	< 0.001

3.8 Undescribed annotated coding space

Table 7: Comparative results of undescribed coding space and one-way ANOVA Tukey HSD post hoc test for undescribed coding space.The differences on example of Prokka vs Bakta means that on average Prokka had 829665.56 bp larger undescribed coding space.

Comparison	Mean difference	Lower end point of CI 95%	Upper end point of CI 95%	adj P-value
Prokka vs Bakta	829665.56	827720.19	831610.93	< 0.001
Prokka vs EggNOG-mapper	690833.26	688887.89	692778.63	< 0.001
Prokka vs PGAP	731815.84	729870.48	733761.21	< 0.001
EggNOG-mapper vs Bakta	138832.30	136886.93	140777.67	< 0.001
PGAP vs Bakta	97849.72	95904.35	99795.08	< 0.001
PGAP vs EggNOG-mapper	-40982.58	-42927.95	-39037.21	< 0.001

3.9 Differences in annotating assembly gaps between tools.

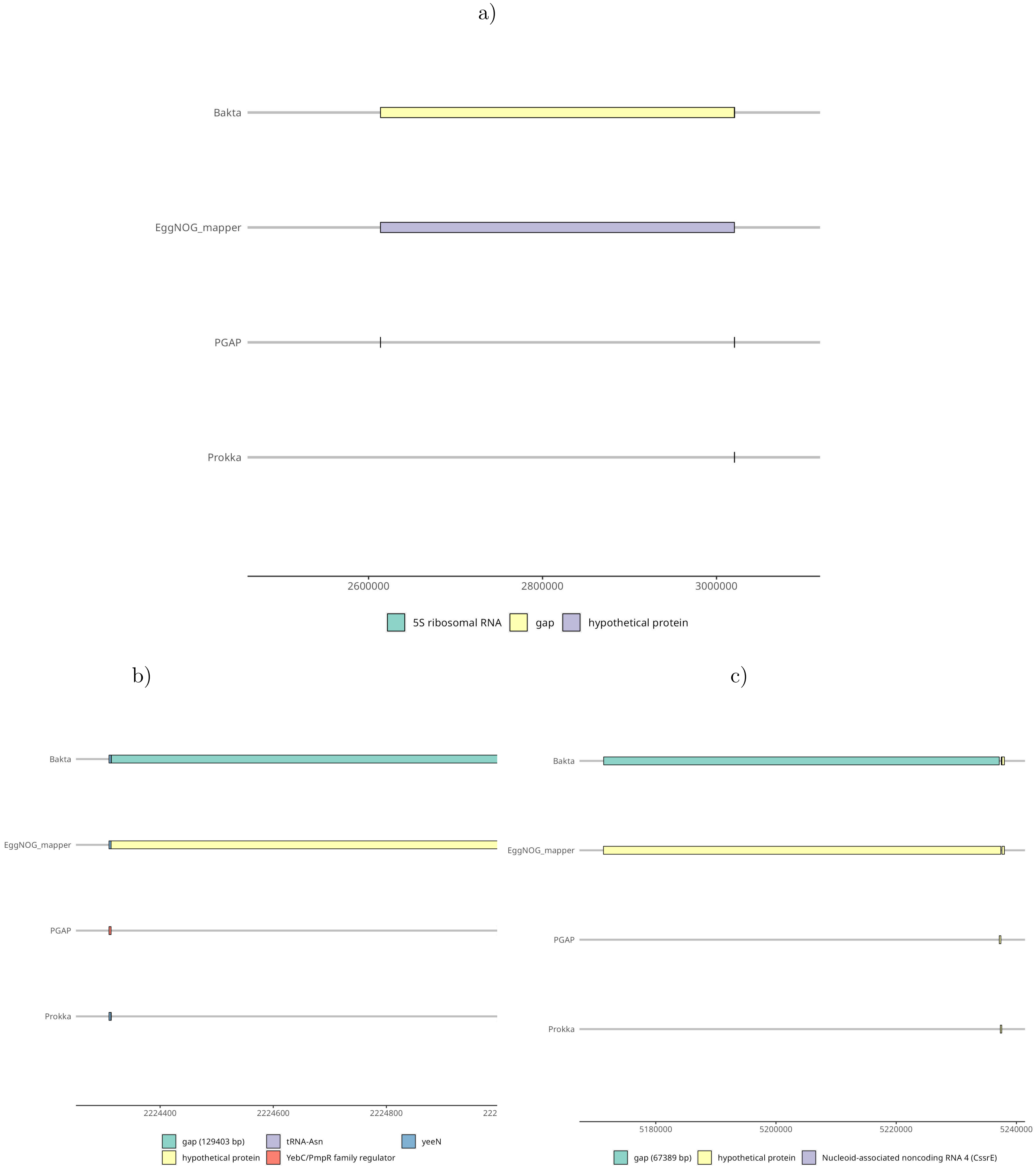


Figure 5: Examples of unusually long hypothetical proteins identified within EggNOG-mapper on the examples found in GCA_014216915.1. Specific genomic regions with found anomalies are: (a) 2490000 to 3090000 bp, (b) 2224285 to 2355669 bp, (c) 5170000 to 5239000 bp.

3.10 Impact on rRNA and tRNA discoverability in *E. coli*

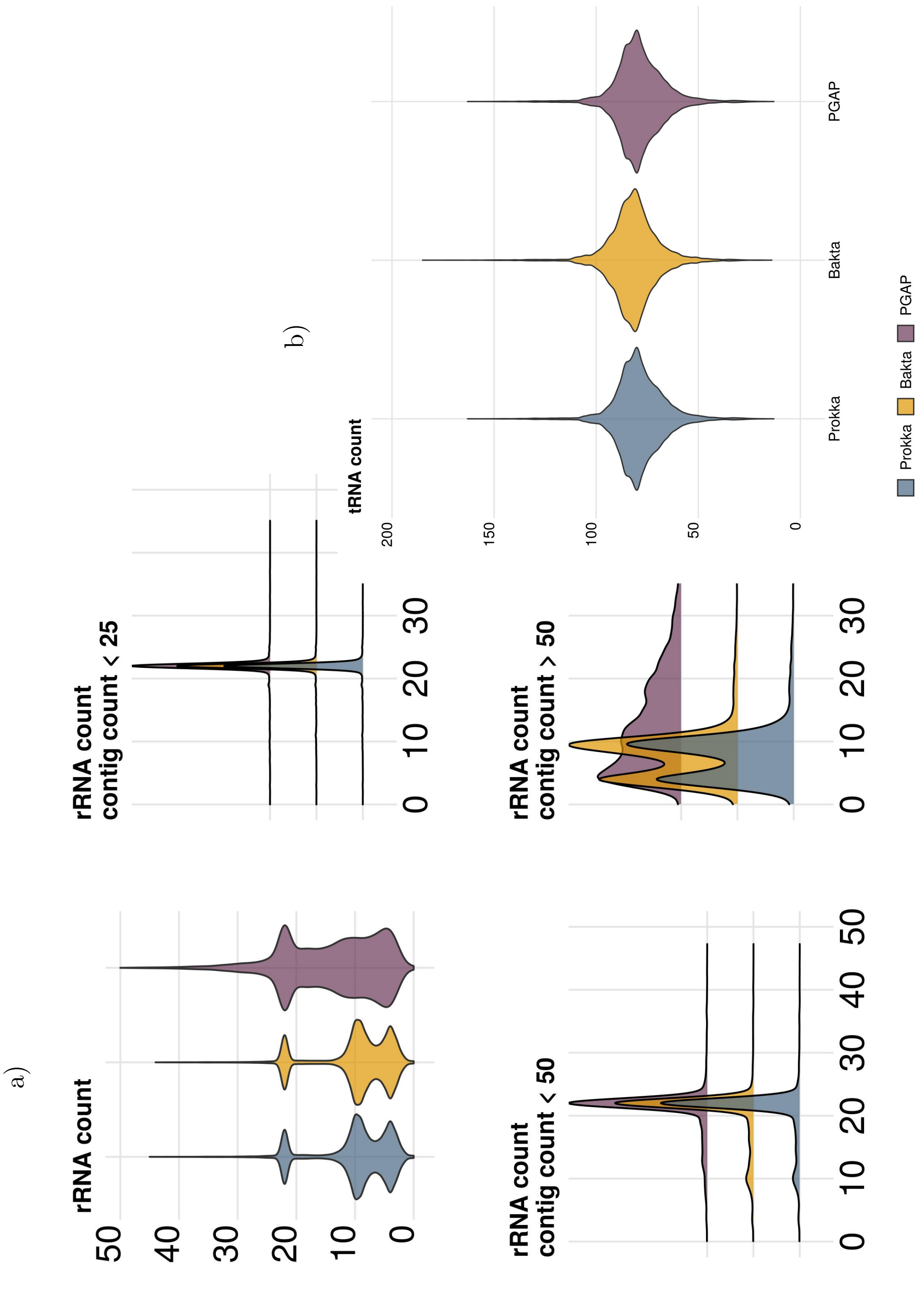


Figure 6: Analysis of contig counts and their impact on rRNA/tRNA discoverability in *Escherichia coli*. a) rRNA identification across contig count bins, with a ridge plot showing rRNA distribution in three groups: <25, <50, and >50 contigs. b) tRNA count distribution.

4 Bacteria kingdom

4.1 Data description

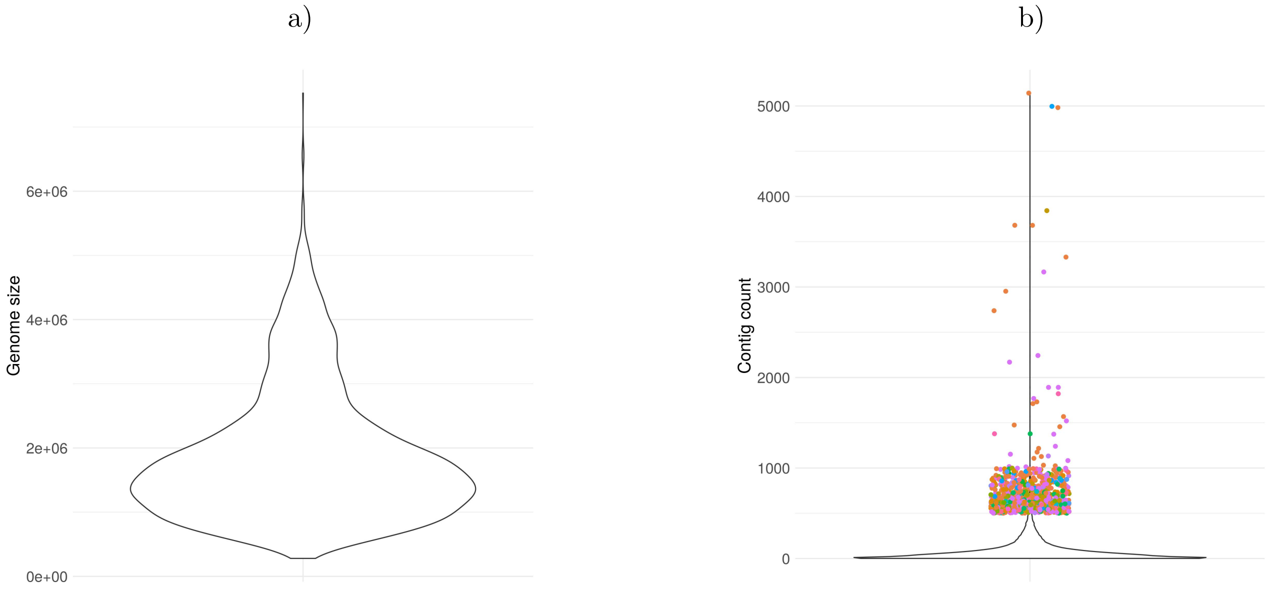


Figure 7: Basic metrics characterizing the Bacteria kingdom group within the analyzed samples. Each panel presents a distinct attribute: (A) Genome size derived from genome_size variable, (B) Contig count derived from contig_count variable. Major outliers in the respective categorise are shown as a dots, colour coded by their phylia obtained from gtdb_taxonomy variable.

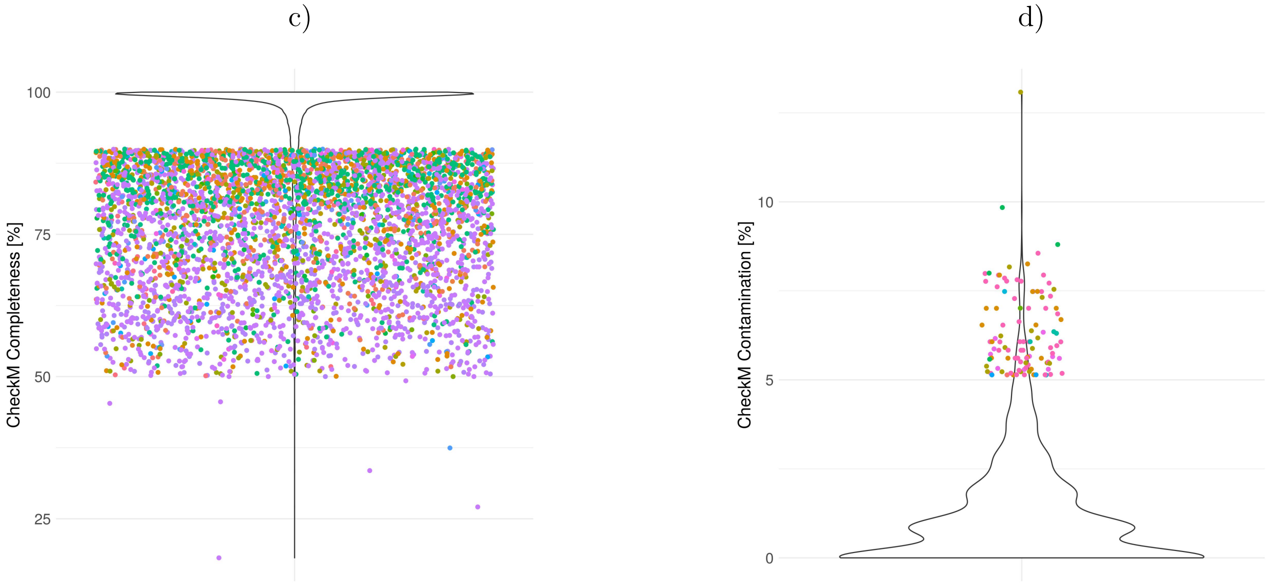


Figure 8: Basic metrics characterizing the Bacteria kingdom group within the analyzed samples. Each panel presents a distinct: (C) Genome completeness derived from GTDB checkm_completeness variable, and (D) Genome contamination derived from GTDB checkm_contamination variable. Major outliers in the respective categorise are shown as a dots, colour coded by their phylia obtained from gtdb_taxonomy variable.

4.2 Taxonomy

4.2.1 Total coding length

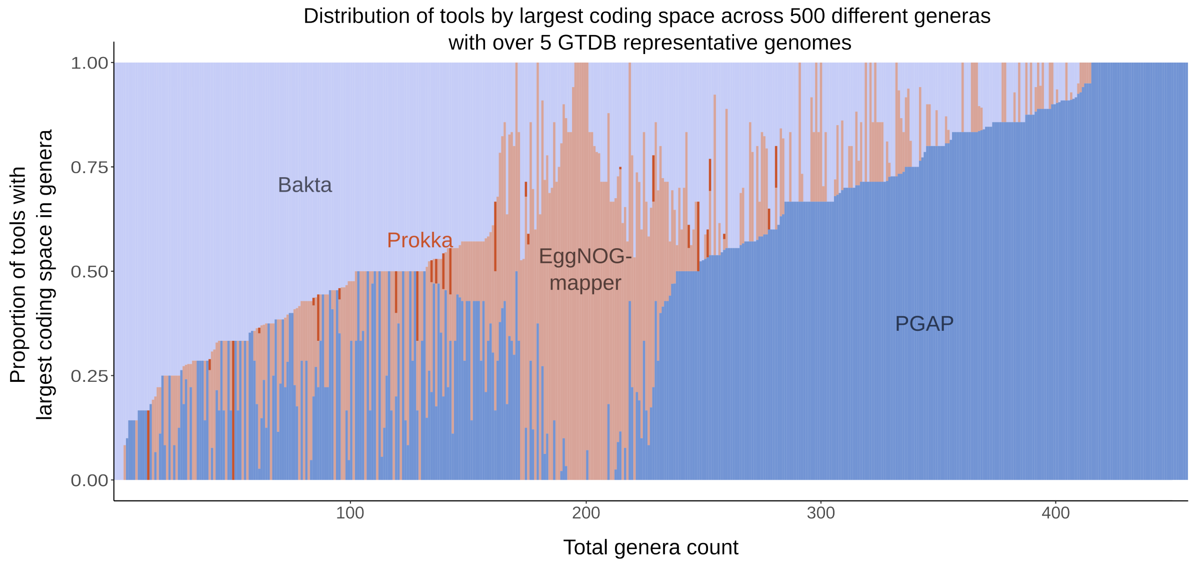


Figure 9: Competitive tool performance for the largest total coding space per species grouped by genus. The barcharts highlights the number of times a tool has achieved the best performance in maximizing total coding space. The difference in performance between the best and second-best tools was not considered by this approach.

4.2.2 Feature length comparison by phyla

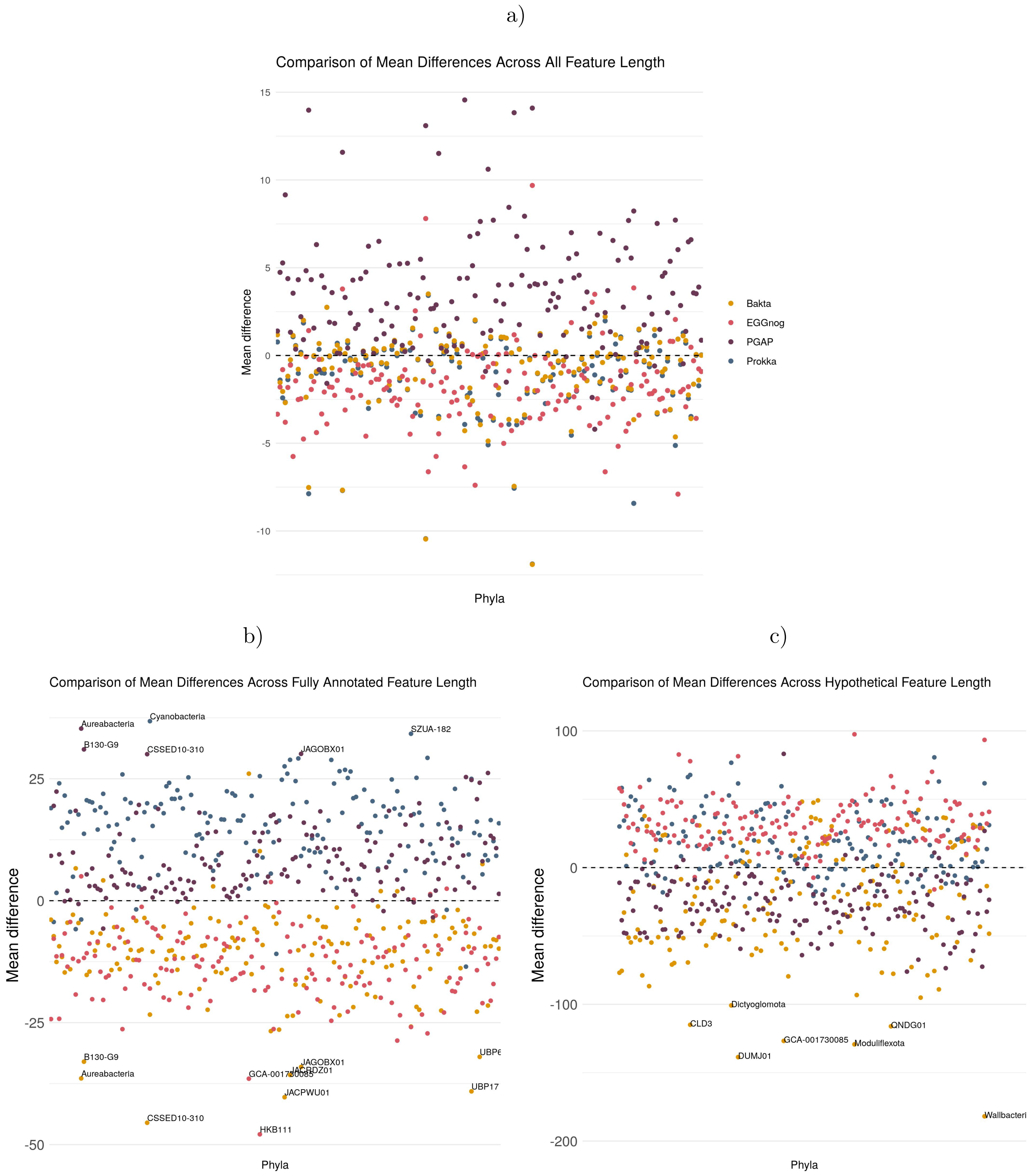


Figure 10: Comparison of mean differences in feature lengths for Prokka, Bakta, EggNOG-mapper, and PGAP across various phyla. Dashed horizontal lines at y = 0 indicate no mean difference. Subplot A) shows differences in mean lengths of all features (including hypothetical and fully annotated), with phylum names annotated where the absolute difference exceeds 30. B) displays differences in mean lengths of described features, with annotations for absolute differences over 30. C) illustrates differences in mean lengths of undescribed features, with annotations for absolute differences over 100.

4.3 Analysis of rRNA prediction variance

Table 8: Top 20 samples with highest count of rRNA detected by Prokka

id	Prokka rRNA	Bakta rRNA	PGAP rRNA	Contig count
GCA_003674045.1	134	65	143	227
GCA_002162355.1	108	108	108	1
GCA_001623875.1	85	82	84	25
GCA_002243515.1	80	80	80	1
GCA_900176885.1	67	65	64	41
GCA_900291985.1	59	55	54	16
GCA_004006295.1	54	54	54	2
GCA_008931805.1	53	53	53	2
GCA_002355375.1	51	51	51	1
GCA_013248975.1	51	51	51	3
GCA_900460535.1	50	49	50	5
GCA_900465055.1	50	50	50	1
GCA_000196255.1	49	49	49	3
GCA_002019605.1	49	49	49	17
GCA_002019645.1	49	49	48	8
GCA_900002575.1	49	48	49	1
GCA_002109385.1	47	47	47	1
GCA_016811915.1	47	47	47	1
GCA_014879975.1	46	46	46	2
GCA_900199725.1	47	46	40	69

Table 9: Top 20 samples with highest count of rRNA detected by Bakta

id	Prokka rRNA	Bakta rRNA	PGAP rRNA	Contig count
GCA_002162355.1	108	108	108	1
GCA_001623875.1	85	82	84	25
GCA_002243515.1	80	80	80	1
GCA_003674045.1	134	65	143	227
GCA_900176885.1	67	65	64	41
GCA_900291985.1	59	55	54	16
GCA_004006295.1	54	54	54	2
GCA_008931805.1	53	53	53	2
GCA_002355375.1	51	51	51	1
GCA_013248975.1	51	51	51	3
GCA_900465055.1	50	50	50	1
GCA_000196255.1	49	49	49	3
GCA_002019605.1	49	49	49	17
GCA_002019645.1	49	49	48	8
GCA_900460535.1	50	49	50	5
GCA_900002575.1	49	48	49	1
GCA_002109385.1	47	47	47	1
GCA_016811915.1	47	47	47	1
GCA_014879975.1	46	46	46	2
GCA_900199725.1	47	46	40	69

Table 10: Top 20 samples with highest count of rRNA detected by PGAP

id	Prokka rRNA	Bakta rRNA	PGAP rRNA	Contig count
GCA_000415505.1	22	17	227	1731
GCA_003674045.1	134	65	143	227
GCA_002162355.1	108	108	108	1
GCA_009720735.1	32	28	93	151
GCA_014502795.1	18	18	90	218
GCA_016587775.1	14	10	90	730
GCA_001623875.1	85	82	84	25
GCA_000986785.1	35	31	80	104
GCA_002243515.1	80	80	80	1
GCA_013359935.1	18	18	78	206
GCA_002257705.1	28	21	75	165
GCA_003148565.1	26	26	71	71
GCA_018332455.1	1	19	69	143
GCA_002897295.1	15	19	67	197
GCA_002008345.1	21	21	66	138
GCA_018403325.1	17	18	64	101
GCA_900176885.1	67	65	64	41
GCA_016902295.1	12	12	61	265
GCA_003865095.1	22	22	58	178
GCA_018333315.1	12	11	56	146

4.3.1 Capability of PGAP to annotate rRNA features across high contig genomes

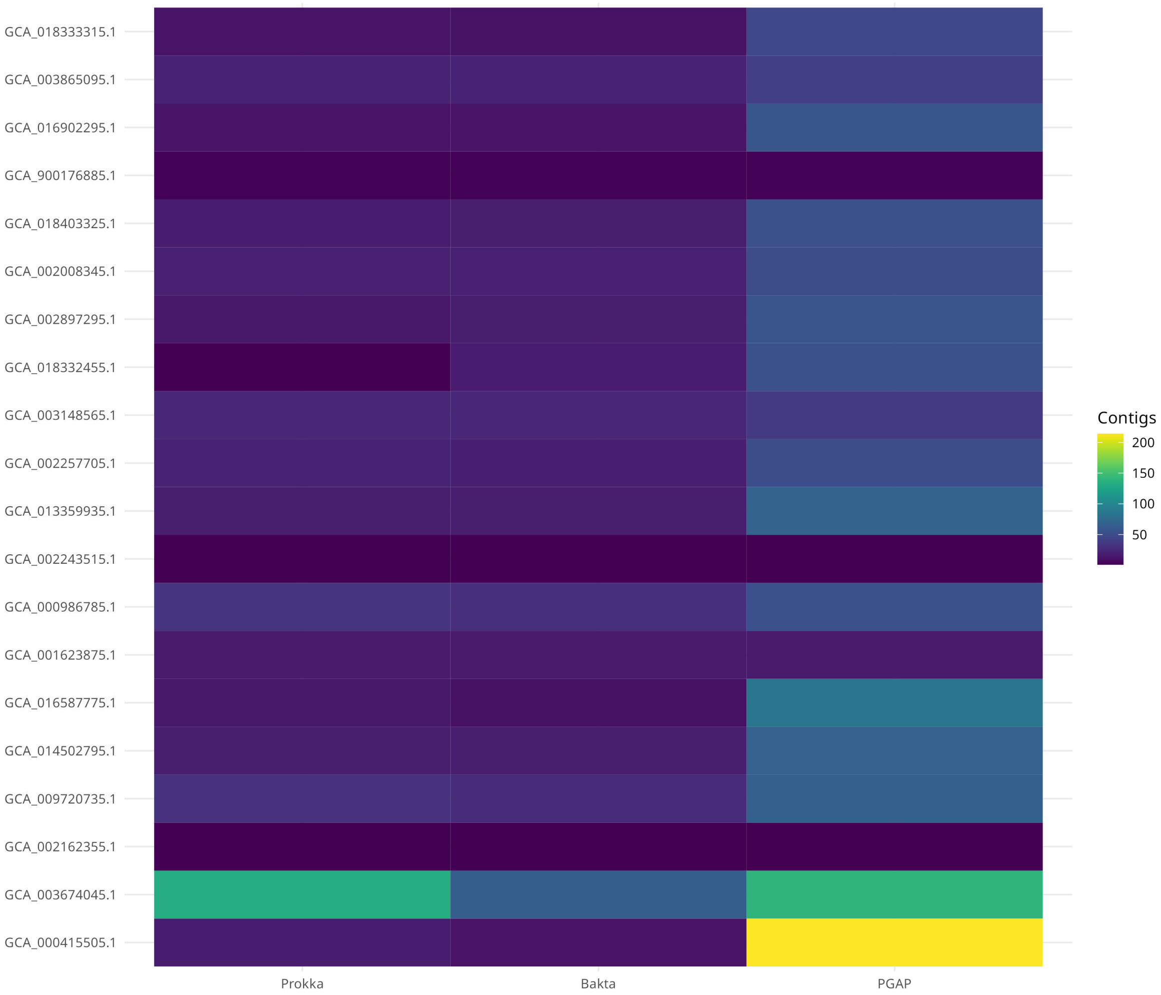


Figure 11: Heatmap of the abundance of unique contigs with rRNA features in the 20 samples with the highest rRNA counts. Color intensity within each cell represents the relative abundance of unique contigs identified as containing rRNA sequences. The raw counts and total contig counts are displayed in Supplementary Tab. 10.

5 Archaea

5.1 Data description

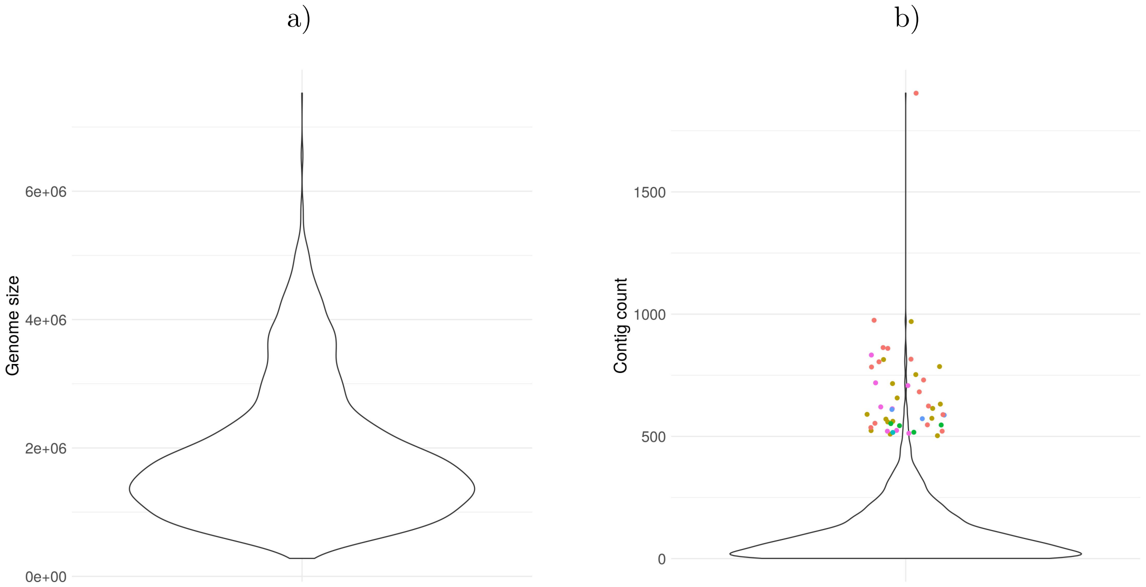


Figure 12: Basic metrics characterizing the Archaea subgroup within the analyzed samples. Each panel presents a distinct attribute: (A) Genome size derived from genome_size variable, (B) Contig count derived from contig_count variable. Major outliers in the respective categorise are shown as a dots, colour coded by their phylia obtained from gtdb_taxonomy variable.

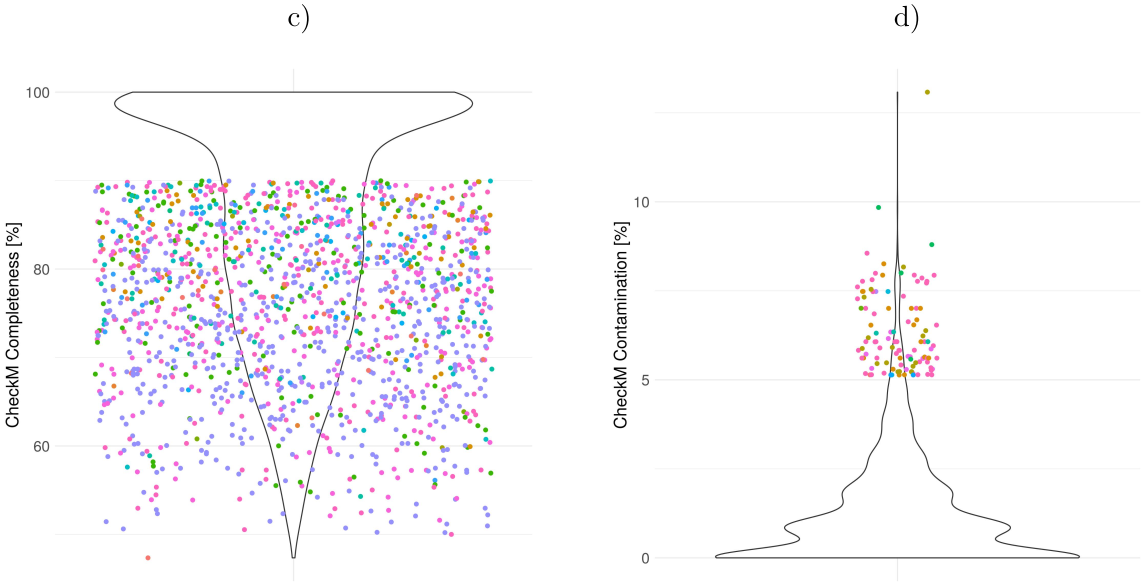


Figure 13: Basic metrics characterizing the Archaea subgroup within the analyzed samples. Each panel presents a distinct: (C) Genome completeness derived from GTDB checkm_completeness variable, and (D) Genome contamination derived from GTDB checkm_contamination variable. Major outliers in the respective categorise are shown as a dots, colour coded by their phylia obtained from gtdb_taxonomy variable.

5.2 RNA comparison for Archaea subgroup

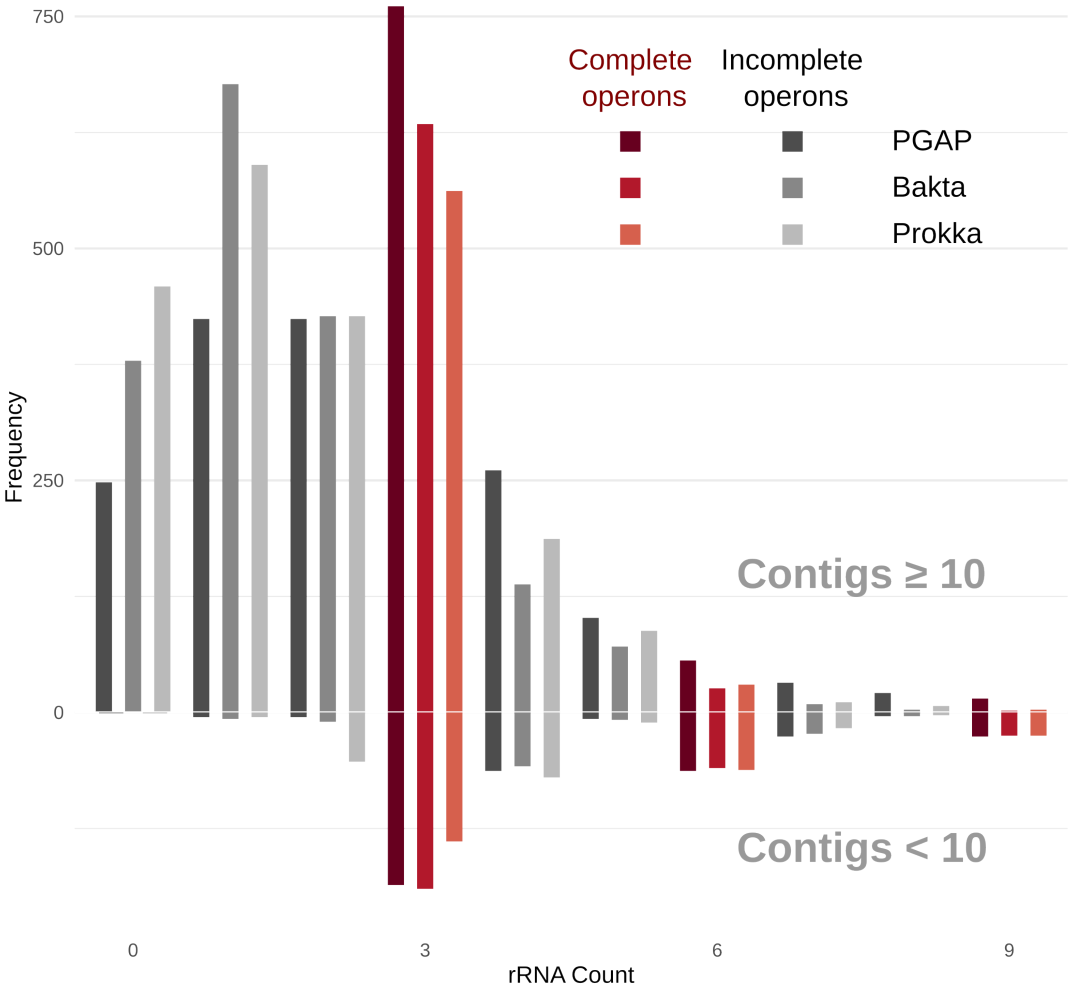
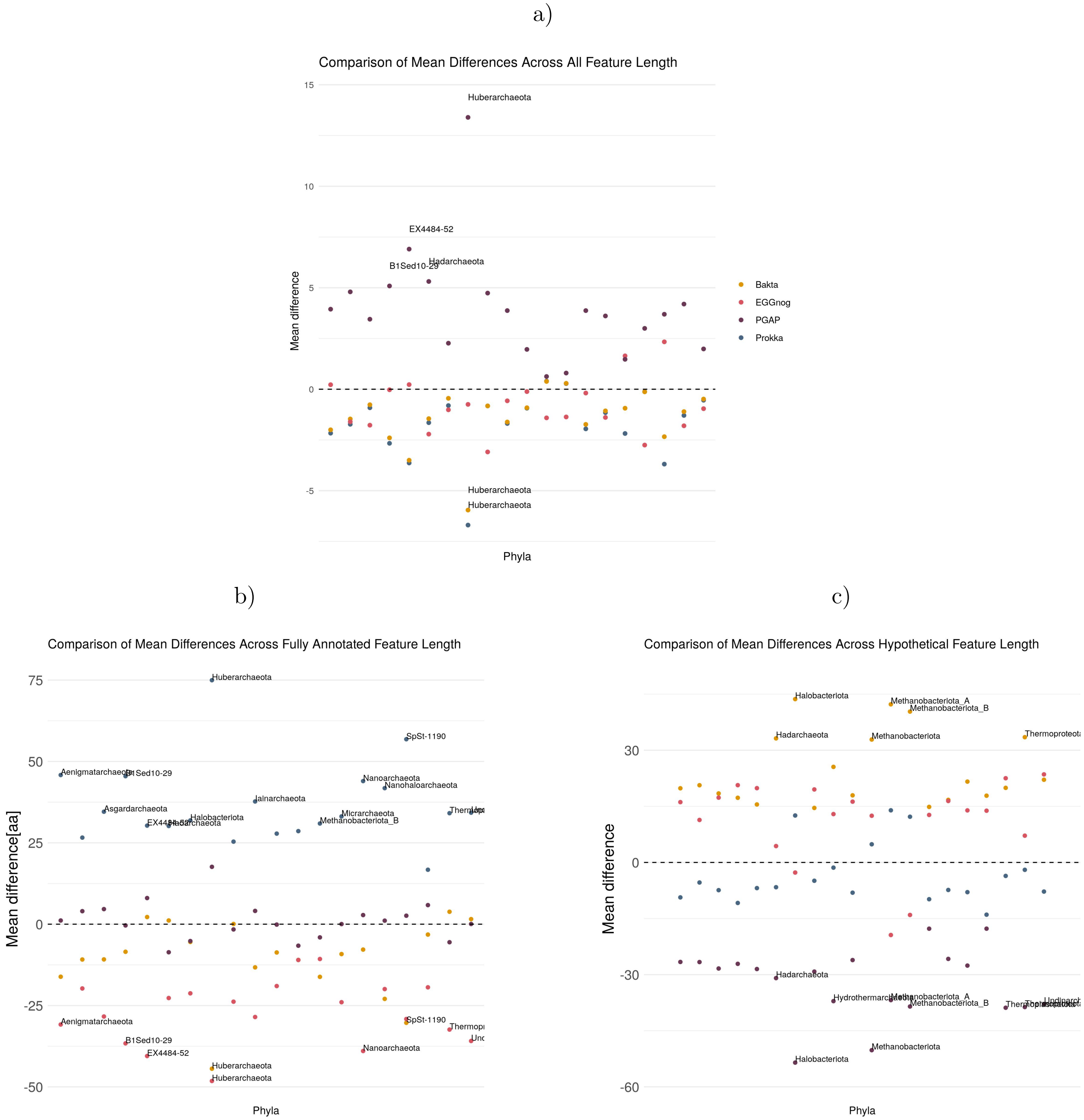


Figure 14: Distribution of rRNA counts in complete and incomplete operons across each tool in Archaea. The upper panel displays the more fragmented genomes, characterized by having more than 10 contigs. In contrast, the bottom panel shows the less fragmented samples, with fewer than 10 contigs. Complete operons, containing multiples of 3 rRNA genes, are highlighted in red, while incomplete operons are displayed in gray.

5.3 Feature length comparison by phyla


a)

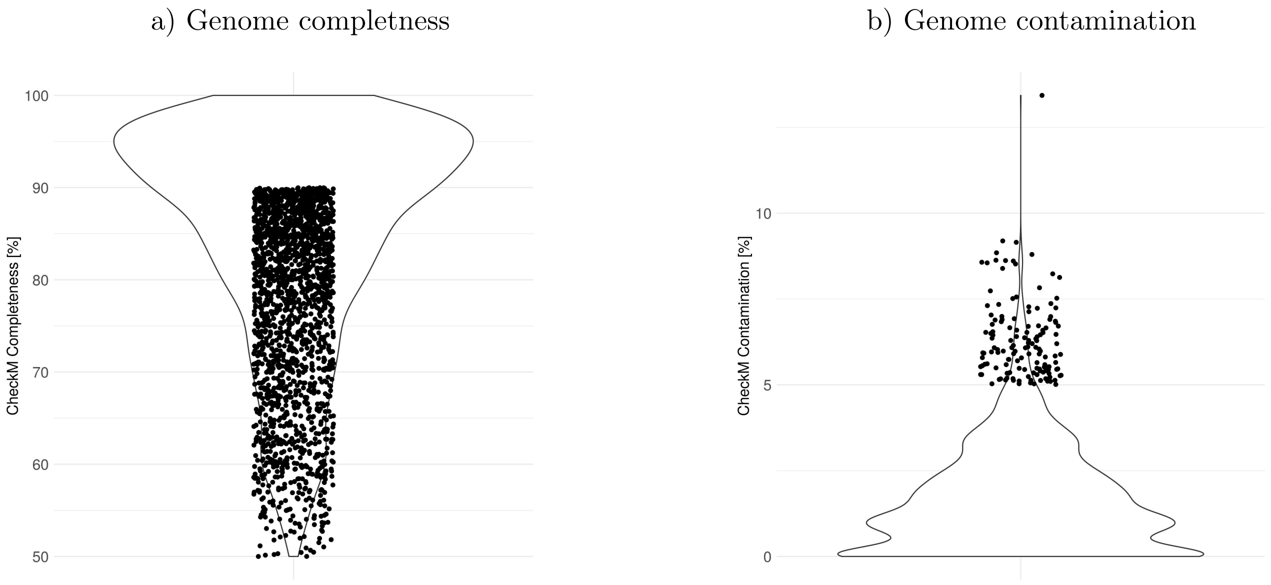
b)

c)

Figure 15: Comparison of mean differences across feature lengths between tools in phyla. Dashed horizontal lines at y = 0, denotes no difference in means. a) Difference between total mean length (including both hypothetical and fully annotated features), phylum name is annotated in cases where absolute difference from the mean of all tools is higher than 30. b) Difference in described mean lenght, phylum name is annotated in cases where absolute differenceis higher than 30, c) Difference in undescribed mean lenght, phylium name is annotated in cases where absolute difference is higher than 100.

6 Metagenome-assambled genomes (MAGs)

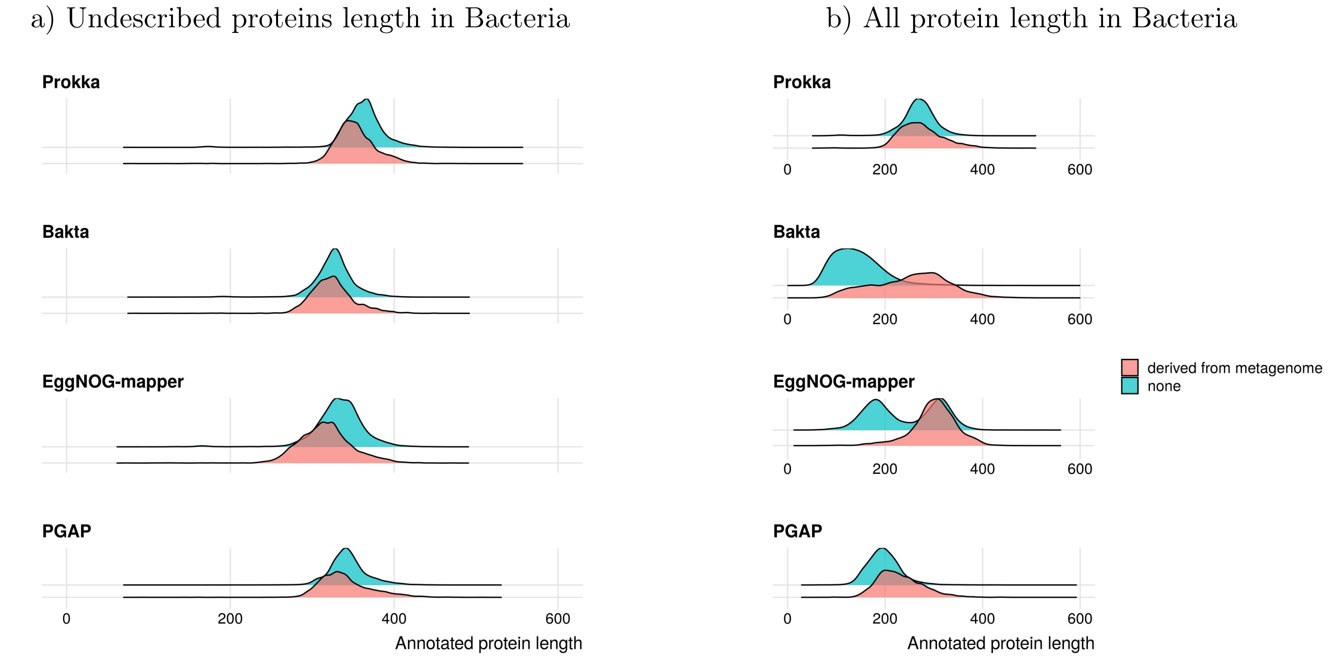
6.1 Data description


a) Genome completness

b) Genome contamination

Figure 16: Basic metrics characterizing the MAGs subgroup within the analyzed samples. Each panel presents a distinct attribute reported in GTDB/NCBI dataset: (a) Genome completeness derived from GTDB checkm_completeness variable, and (b) Genome contamination derived from GTDB checkm_contamination variable.

6.2 Coding length comparison


a) Undescribed proteins length in Bacteria

b) All protein length in Bacteria

Figure 17: Comparison of metagenome protein lengths across analyzed annotation tools in Bacteria. Plot (a) describes the distribution of hypothetical protein lengths, while plot (b) includes all protein lengths.

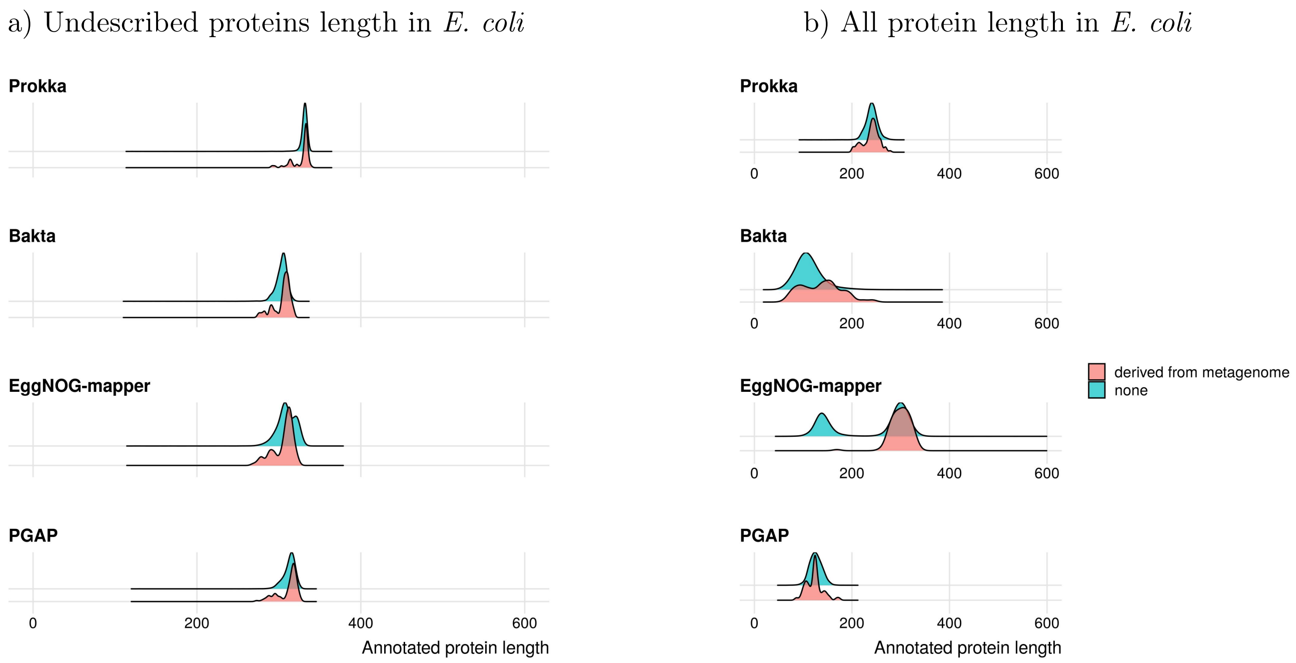


Figure 18: Comparison of metagenome protein lengths across analyzed annotation tools in *E. coli*. Plot (c) describes the distribution of hypothetical protein lengths, while plot (d) includes all protein lengths.

7 Frameshifted genomes

7.1 Genome region comparison

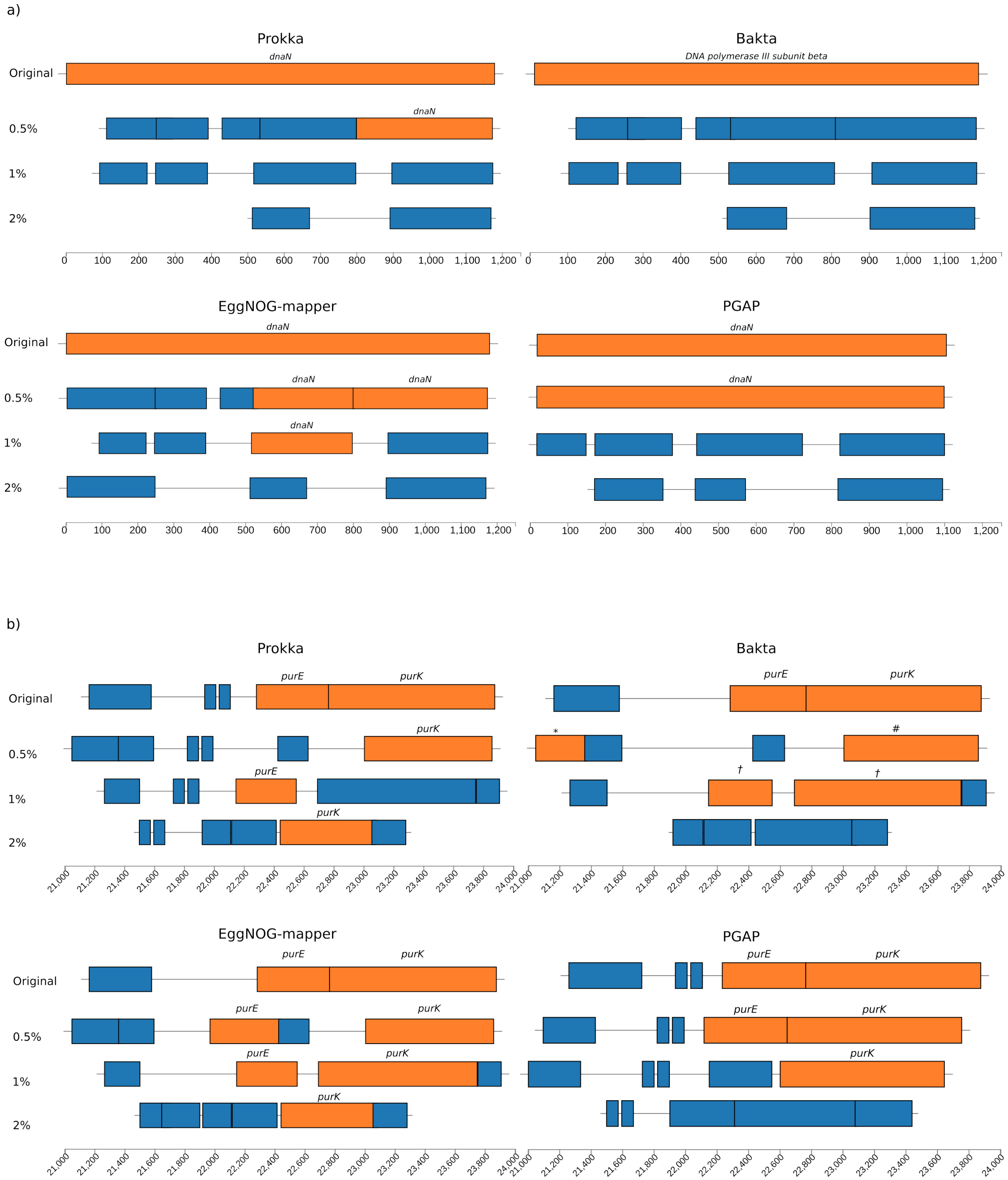


Figure 19: Investigation of gene fragmentation between the original genome sequence (GCA_000008565.1) and sequences with increasing deletion rates (0.5%, 1%, 2% of the full genome). Only gene features fully contained within the investigated regions are displayed. As deletion rates increase, the expected location of features shifts to the left of the plot. Blue features represent undescribed genes, while orange features represent described genes. **a)** Region 1:1,200 bp. **b)** Region 21,000:24,000 bp. Symbols for features with longer names are: * UDG domain-containing gene, # ATP-grasp domain-containing protein, + N5-carboxyaminoimidazole ribonucleotide synthase.

8 Temporal analysis

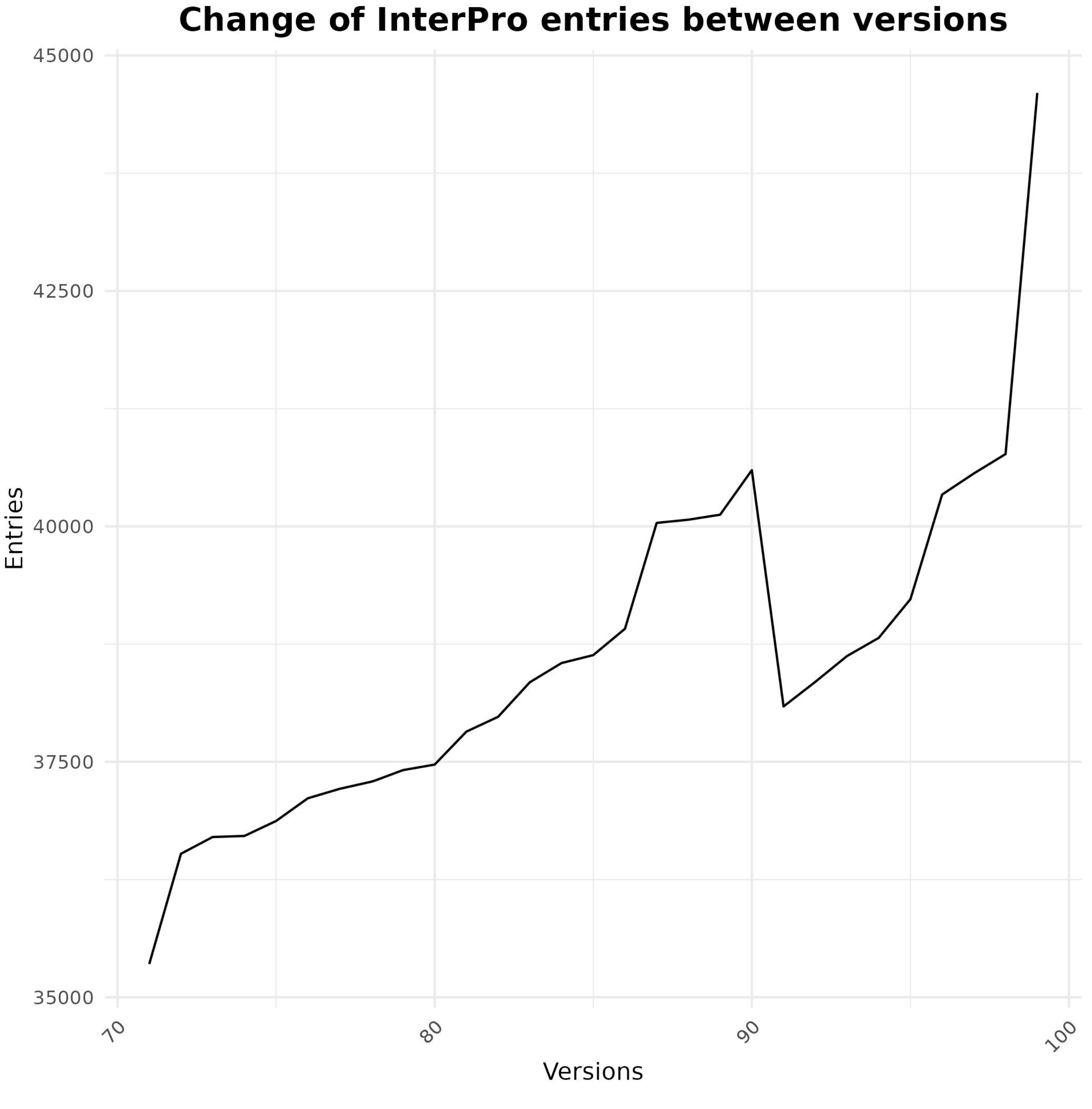
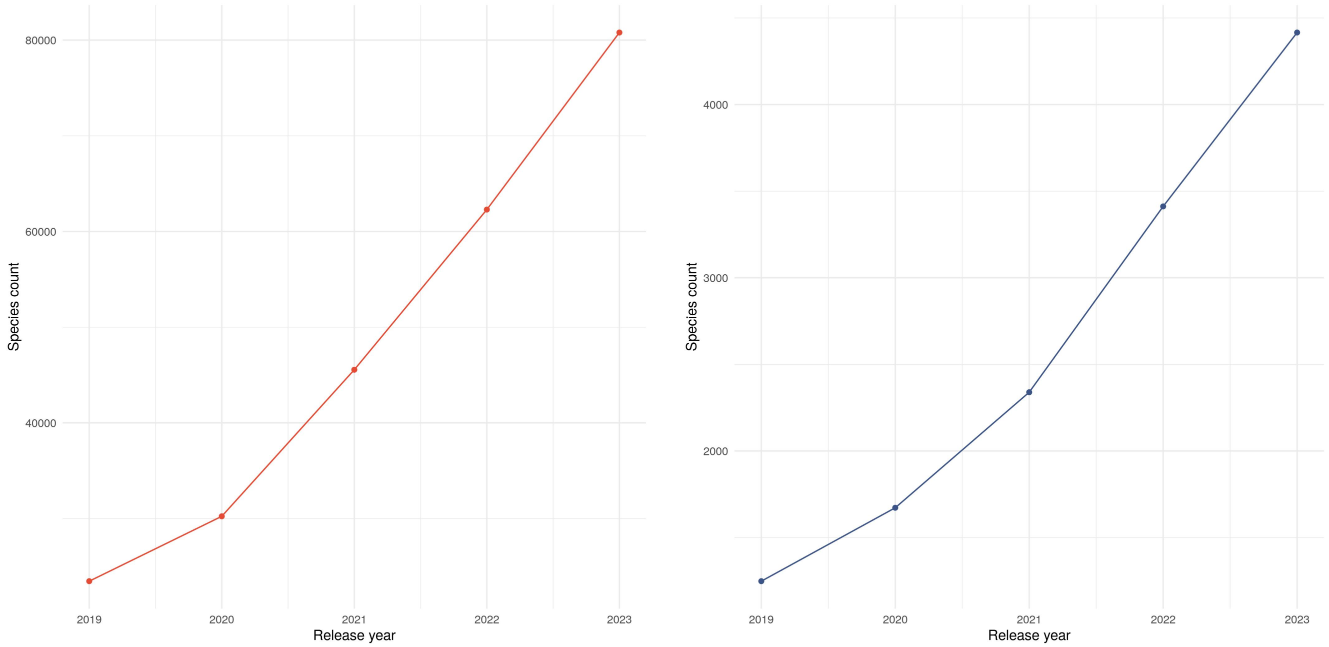


Figure 20: Stability of protein count in anotation results reported by NCBI across reported submission years. For example InterPro witnessed a substantial growth in its entries, expanding from 38,088 to 46,035 within just two years (InterPro 91.0 in October 2022 to 102.0 in October 2024)

8.1 Database increase of unique species across years

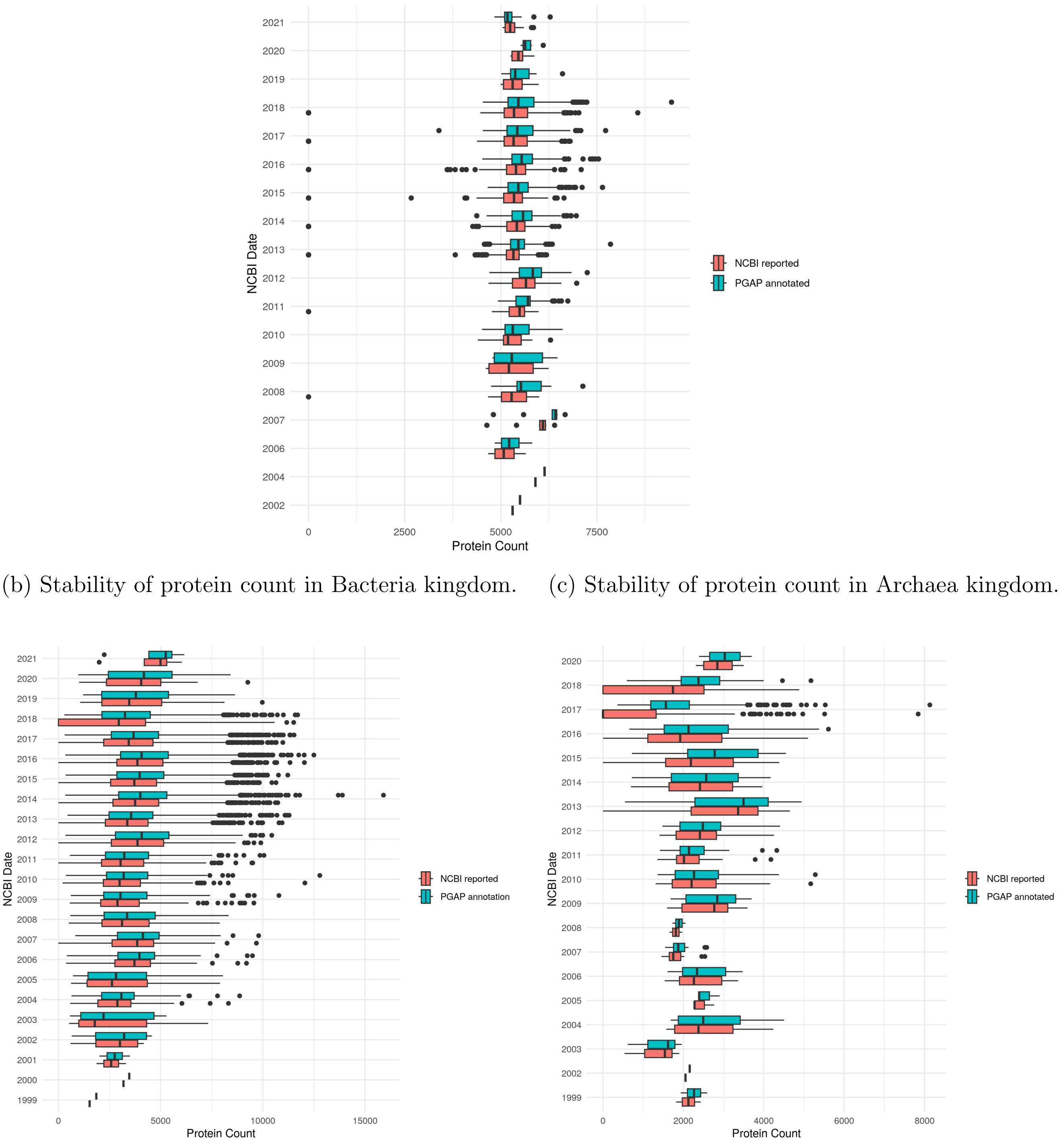


a) GTDB unique Bacteria species increase per GTDB release.

b) GTDB unique Archaea species increase per GTDB release.

Figure 21: Increase in unique species per GTDB release for Bacteria and Archaea.

9 Annotation performance over time



(a) Stability of protein count in *E. coli* samples.

(b) Stability of protein count in Bacteria kingdom.

(c) Stability of protein count in Archaea kingdom.

Figure 22: Stability of reported protein count across submission years for different taxonomic groups. (a) E. coli group. (b) Bacteria kingdom. (c) Archaea kingdom.

10 Workflow visualization

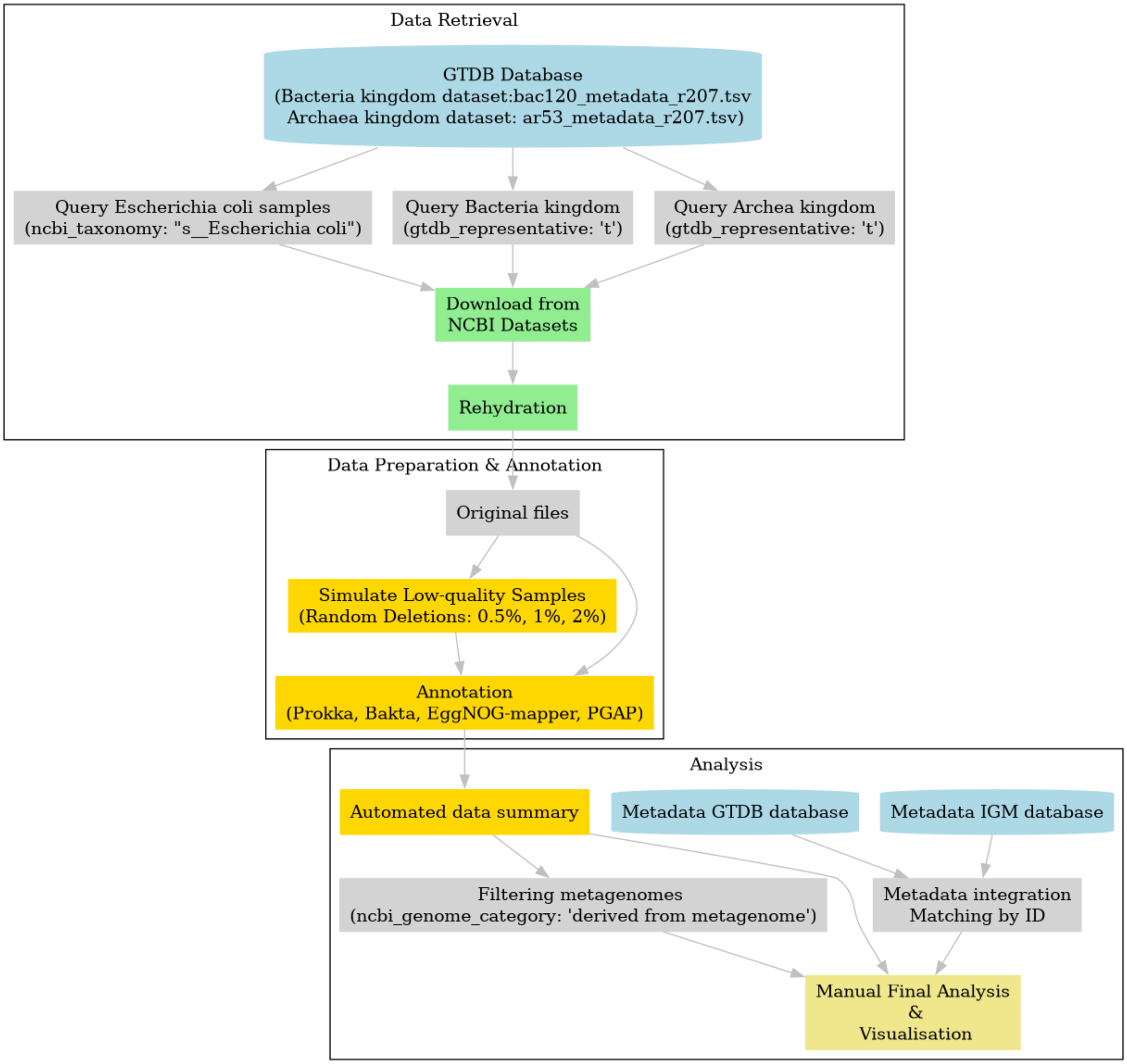

Data Retrieval
GTDB Database
(Bacteria kingdom dataset:bac120_metadata_r207.tsv
Archaea kingdom dataset: ar53_metadata_r207.tsv)
Query Escherichia coli samples
(ncbi_taxonomy: "s__Escherichia coli")
Query Bacteria kingdom
(gtdb_representative: 't')
Query Archea kingdom
(gtdb_representative: 't')
Download from
NCBI Datasets
Rehydration
Data Preparation & Annotation
Original files
Simulate Low-quality Samples
(Random Deletions: 0.5%, 1%, 2%)
Annotation
(Prokka, Bakta, EggNOG-mapper, PGAP)
Analysis
Automated data summary
Metadata GTDB database
Metadata IGM database
Filtering metagenomes
(ncbi_genome_category: 'derived from metagenome')
Metadata integration
Matching by ID
Manual Final Analysis
&
Visualisation

Figure 23: Workflow diagram illustrating the workflow execution. The diagram depicts genome quering and retrieval from the Genome Taxonomy Database (GTDB) and NCBI Datasets, followed by data preparation, annotation using Prokka, Bakta, EggNOG-mapper, and PGAP, automated data summary using RScript, metadata integration, and manual final analysis and visualization. Grey arrows indicate the flow of data and processing steps in the workflow. Color-coded elements highlight different stages: lightblue represents external databases, lightgreen denotes NCBI datatool tool usage, gold indicates steps included in Nextflow workflow, and khaki signifies the final manual analysis.

11 Decision tree

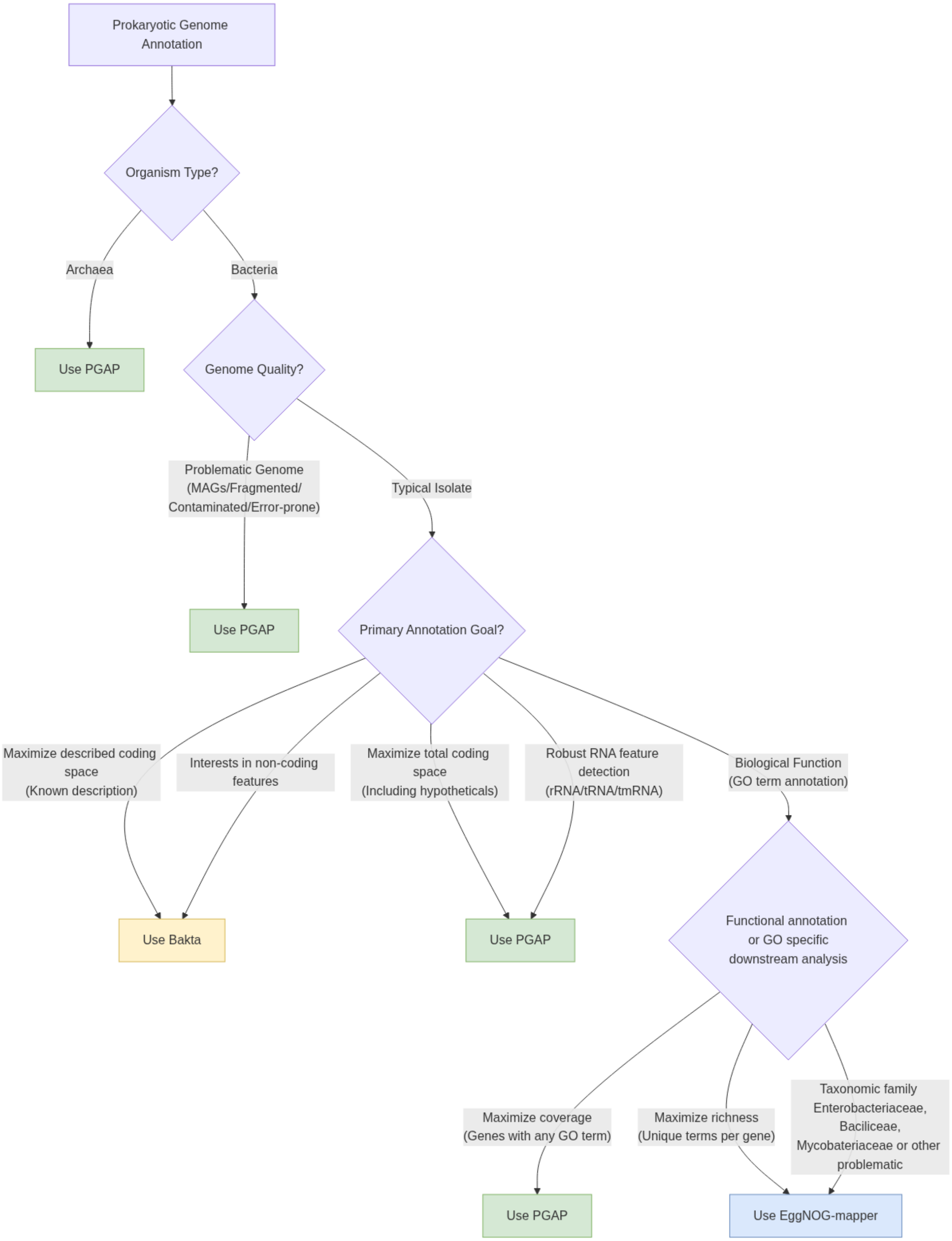

Prokaryotic Genome Annotation
Organism Type?
Archaea
Bacteria
Use PGAP
Genome Quality?
Problematic Genome (MAGs/Fragmented/ Contaminated/Error-prone)
Typical Isolate
Use PGAP
Primary Annotation Goal?
Maximize described coding space (Known description)
Interests in non-coding features
Maximize total coding space (Including hypotheticals)
Robust RNA feature detection (rRNA/tRNA/tmRNA)
Biological Function (GO term annotation)
Use Bakta
Use PGAP
Functional annotation or GO specific downstream analysis
Maximize coverage (Genes with any GO term)
Maximize richness (Unique terms per gene)
Taxonomic family Enterobacteriaceae, Baciliceae, Mycobateriaceae or other problematic
Use PGAP
Use EggNOG-mapper

Figure 24: Simplified decision tree for choosing an appropriate annotation strategy