Distinct K-mers Count and Compression Size: Correlations Across Genome Orders

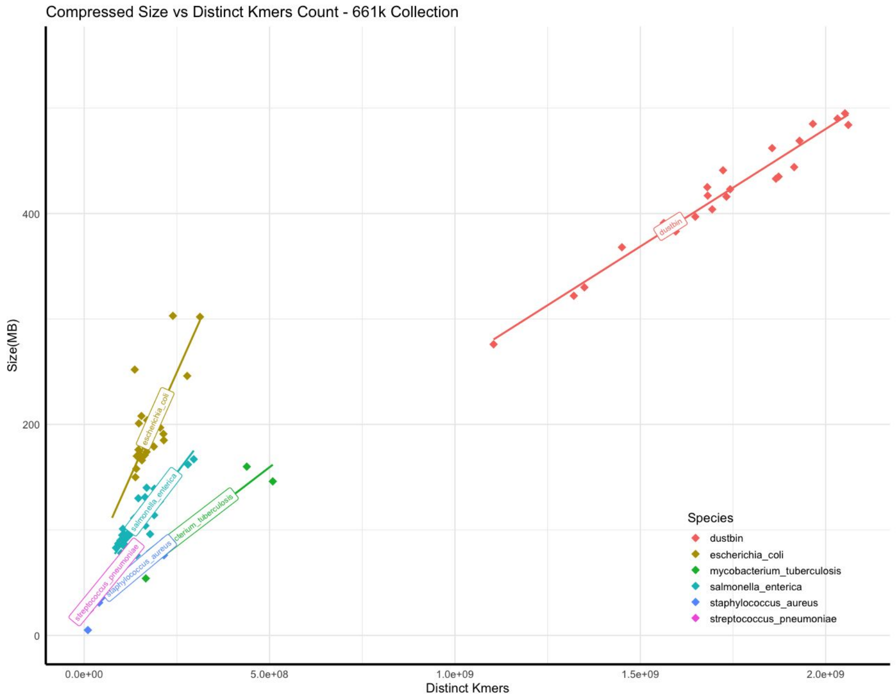
24 Mars 2025

Introduction

We have observed a correlation between distinct k-mers and batch compressed sizes in the miniphy batching of the 661k collection.

The question is:

Will this same phenomenon hold across different genome orderings?"



The Experiment Setup - Data

Data:
The dataset consists of 13 species from the 661k genome collection, which includes more than 10,000 genomes. The metadata is sourced from **BakRep1**.

We sample randomly 10000 genomes from each species.

Total number of genomes from the species:

13 species: 540,545, representing **81% of the collection**. Sampled 130,000, **20% of the collections**

Species included:

Campylobacter jejuni, Clostridioides difficile,Escherichia coli, Klebsiella pneumoniae, Listeria monocytogenes, Listeria monocytogenes B, Mycobacterium tuberculosis, Neisseria meningitidis, Salmonella enterica, Staphylococcus aureus, Streptococcus agalactiae, Streptococcus pneumoniae, Streptococcus pyogenes.

1. Fenske, L., Jelonek, L., Goesmann, A., & Schwengers, O., BakRep – a searchable large-scale web repository for bacterial genomes, Microbial Genomics, 2024.

Orders: Random, Accession, Phylogenetic

For each set of 10,000 genomes of the same species, we reorder them using three methods:

Random: The genomes are randomly shuffled. (bash function shuf)

Accession: The genomes are sorted lexicographically by accession number. (bash function sort)

Phylogenetic: A phylogenetic tree is inferred, and the genomes are ordered based on the leaves from left to right. (attotree[1])

Then we then split the genomes into different size groups.

[1] https://github.com/karel-brinda/attotree

Split the genomes into different size groups

The genomes are then split into groups, starting with a small number of genomes and gradually increasing the number.

For example, with 20 genomes, and 4 groups of increasing size:

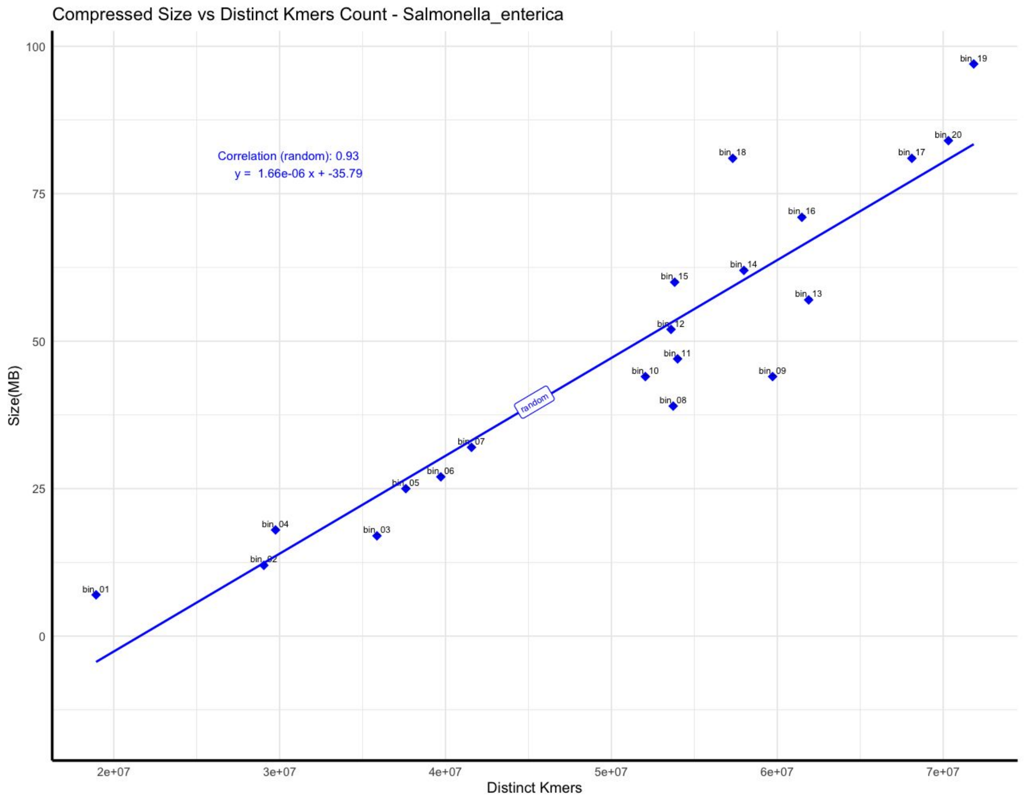
- Subset 1: 2 genomes
- Subset 2: 4 genomes
- Subset 3: 6 genomes
- Subset 4: 8 genomes

In this experiment, we split 10,000 genomes into 20 groups.

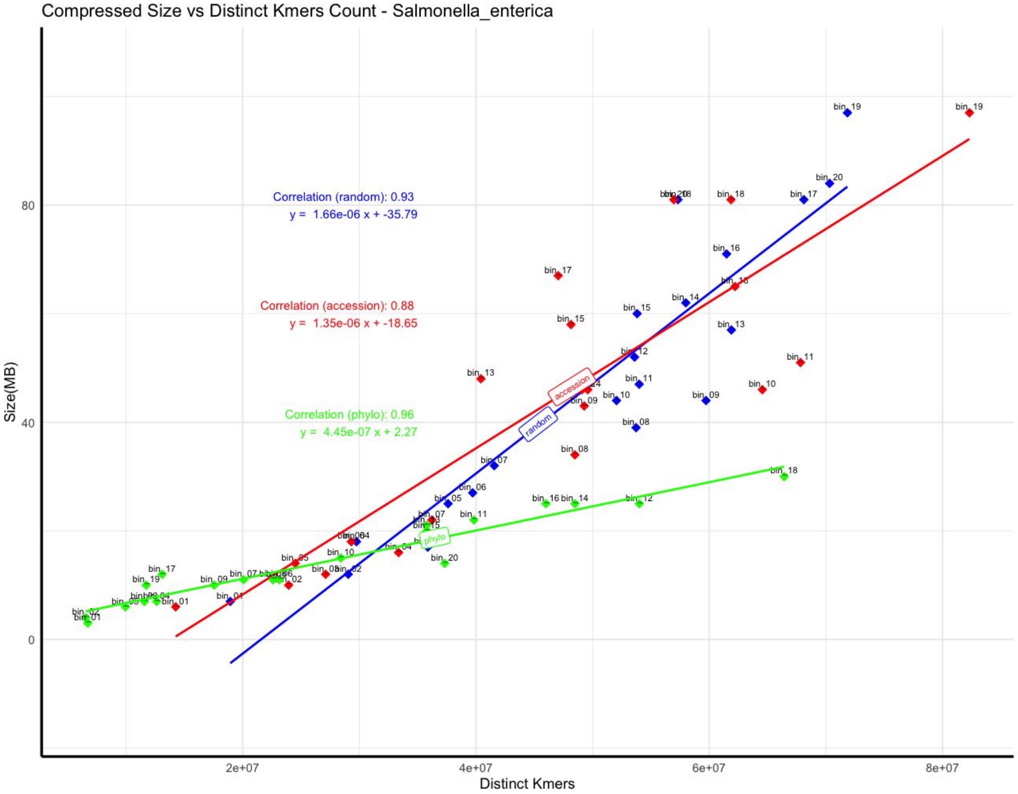
The smallest one has 48, the largest has 952 genomes

Then we estimate the distinct kmers and compress all of them.

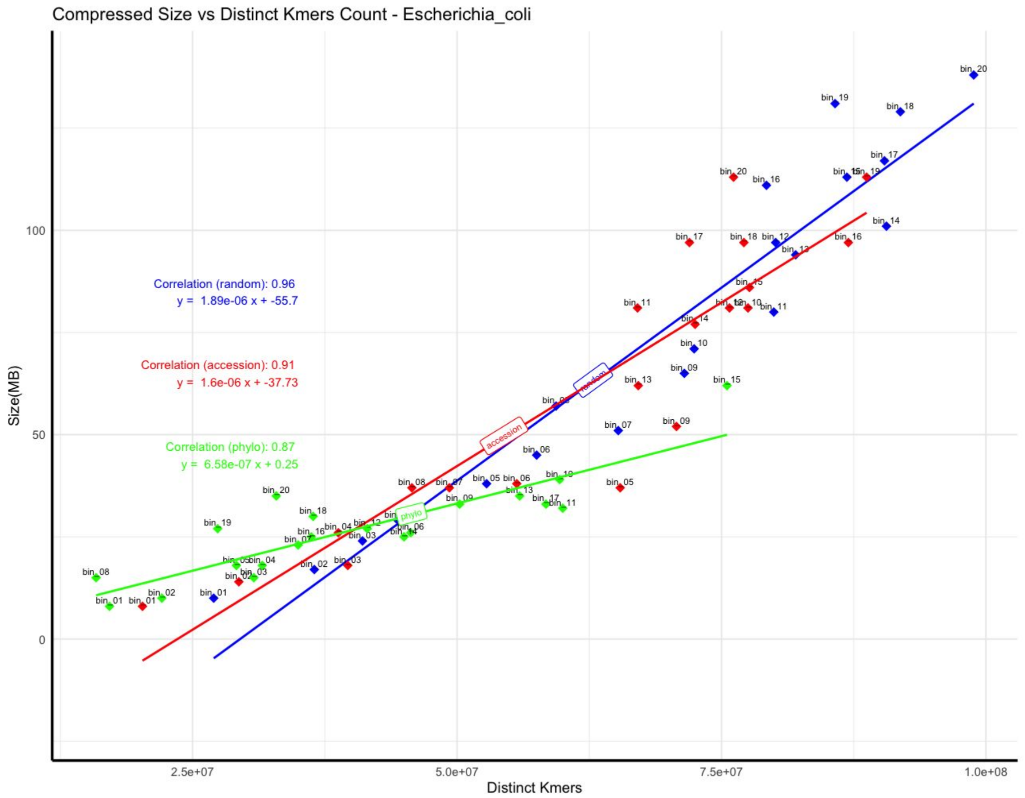
Result: *Salmonella enterica* - random order



Result: *Salmonella enterica* - all orders



Result: *Escherichia coli* - all orders



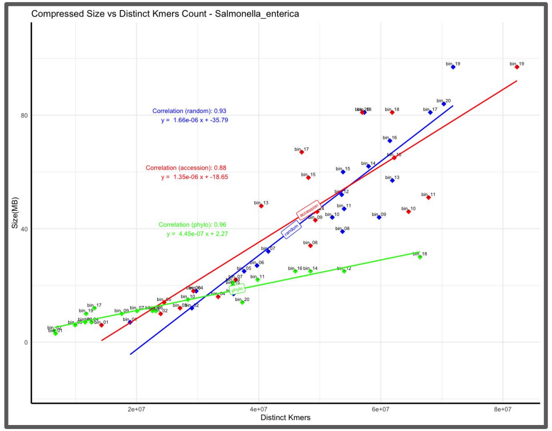
Result: Correlation table for species with more than 10k genomes in 661k

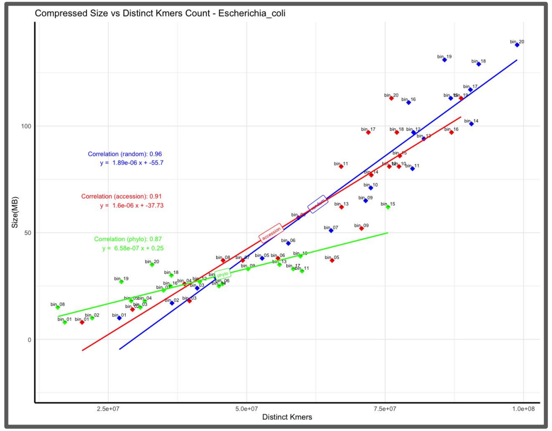
Species	Random	Accession	Phylo
Salmonella enterica	0.93	0.88	0.96
Escherichia coli	0.96	0.81	0.87
Campylobacter D jejuni	0.87	0.6	0.92
Clostridioides difficile	0.82	0.91	0.97
Klebsiella pneumoniae	0.96	0.93	0.95
Listeria monocytogenes	0.95	0.91	0.99
Listeria monocytogenes B	0.87	0.94	0.95
Mycobacterium tuberculosis	0.98	0.99	1
Neisseria meningitidis	0.86	0.82	0.68
Staphylococcus aureus	0.98	0.96	0.93
Streptococcus agalactiae	0.97	0.78	0.96
Streptococcus pneumoniae	0.8	0.82	0.94
Streptococcus pyogenes	0.83	0.63	0.98

Discussion and observation:

Result of Last Week:

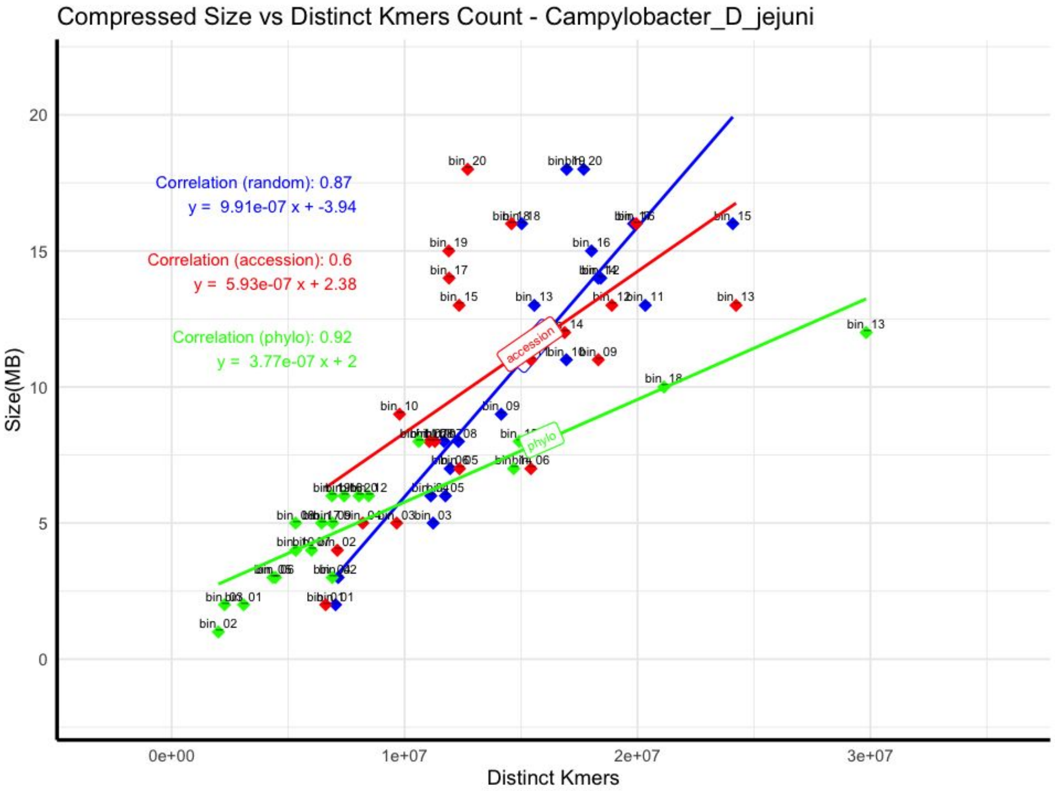
- Developed a comprehensive Snakemake workflow for estimating cardinality and compression size (https://github.com/tmtktmtk/Workspace/tree/main/experiments/019_version_2_snakemake_workflow).
- Tested 3 genome orders on a subset of 13 popular species.
- The correlation between distinct k-mers and compression size holds across different orders and species.
- **Phylogenetic order** performs the best, with the exception of *Neisseria meningitidis*.
- Estimating a phylogenetic tree for 10,000 genomes is slow (~12 hours for 13 species on 14 cores, Mac M4).
- Next step: testing if combining orders (e.g., accession + phylo, random + phylo) would speed up the process while maintaining a good correlation coefficient.
- Upcoming focus: testing the accuracy of the prediction.



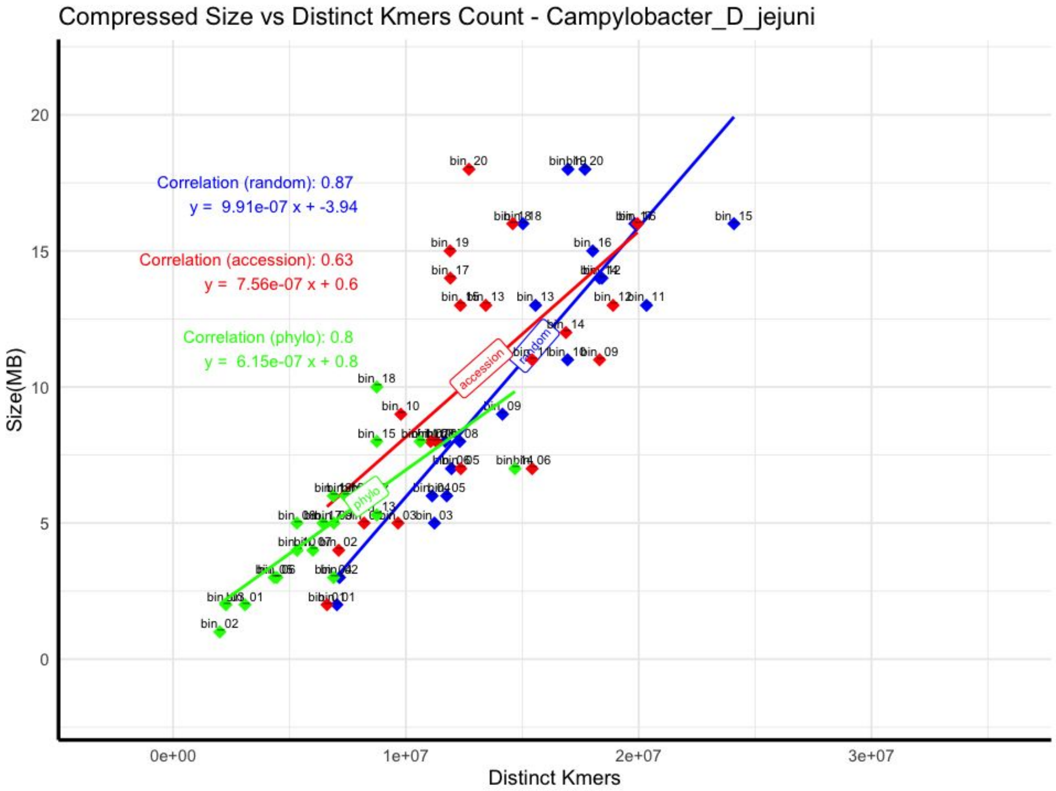


Supplementary

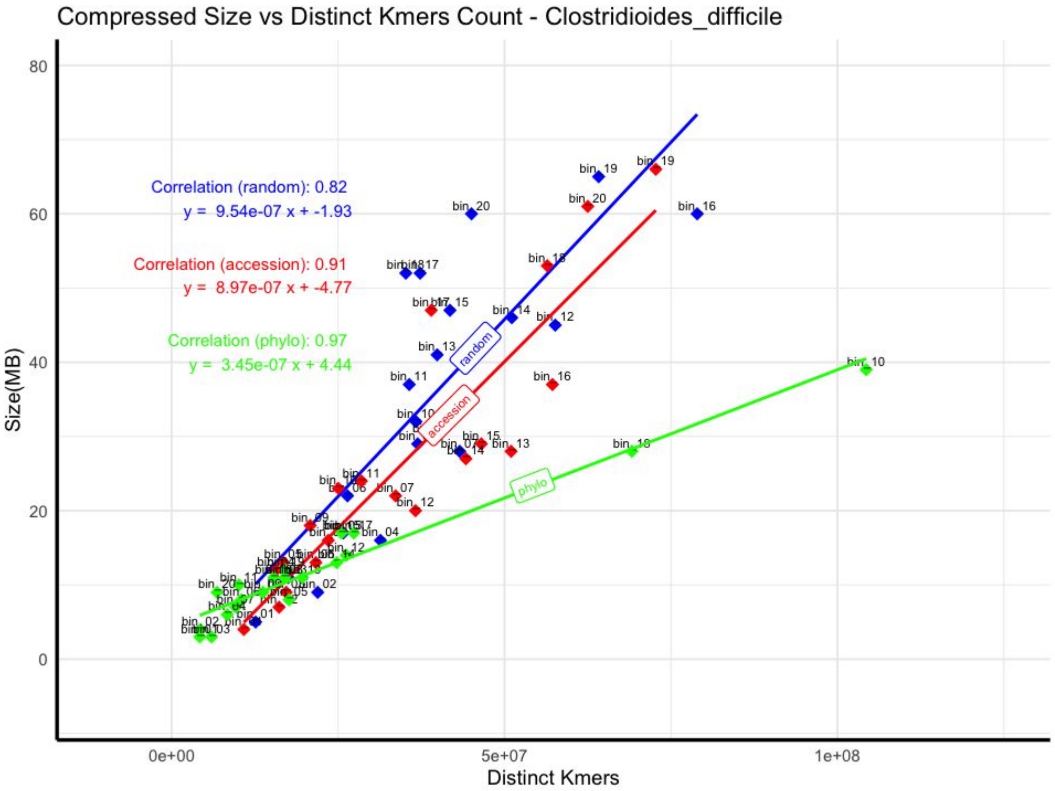
Campylobacter_D_jejuni



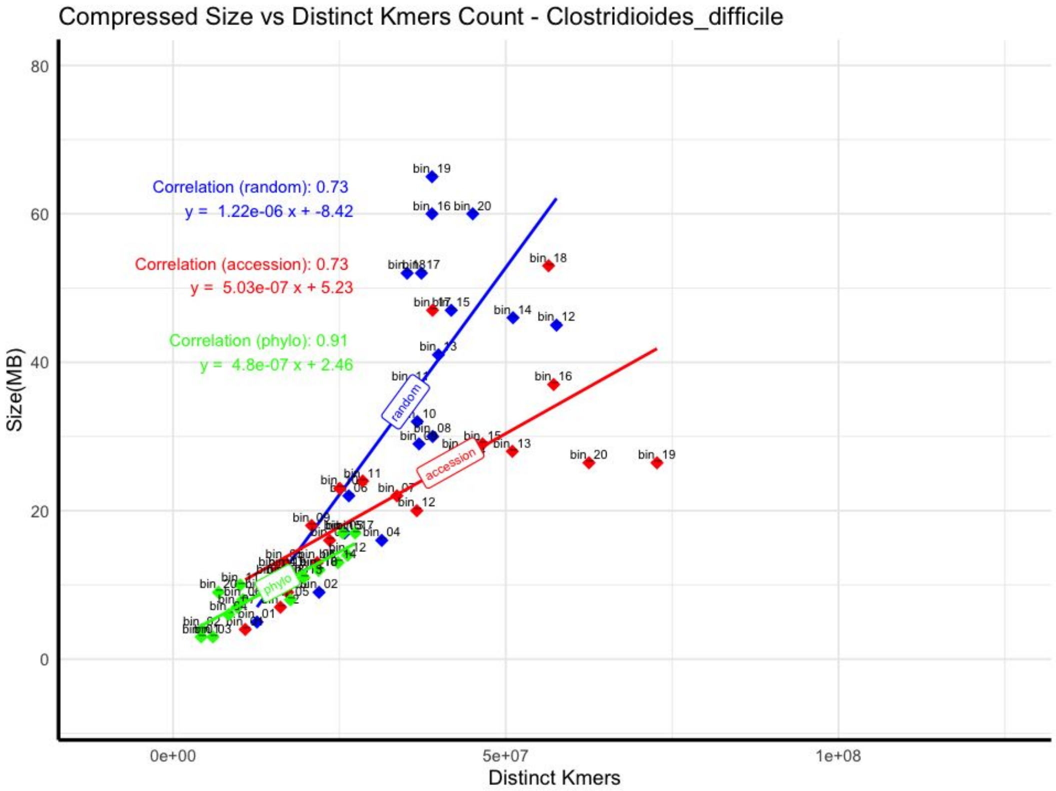
Campylobacter_D_jejuni - replace outliers with means



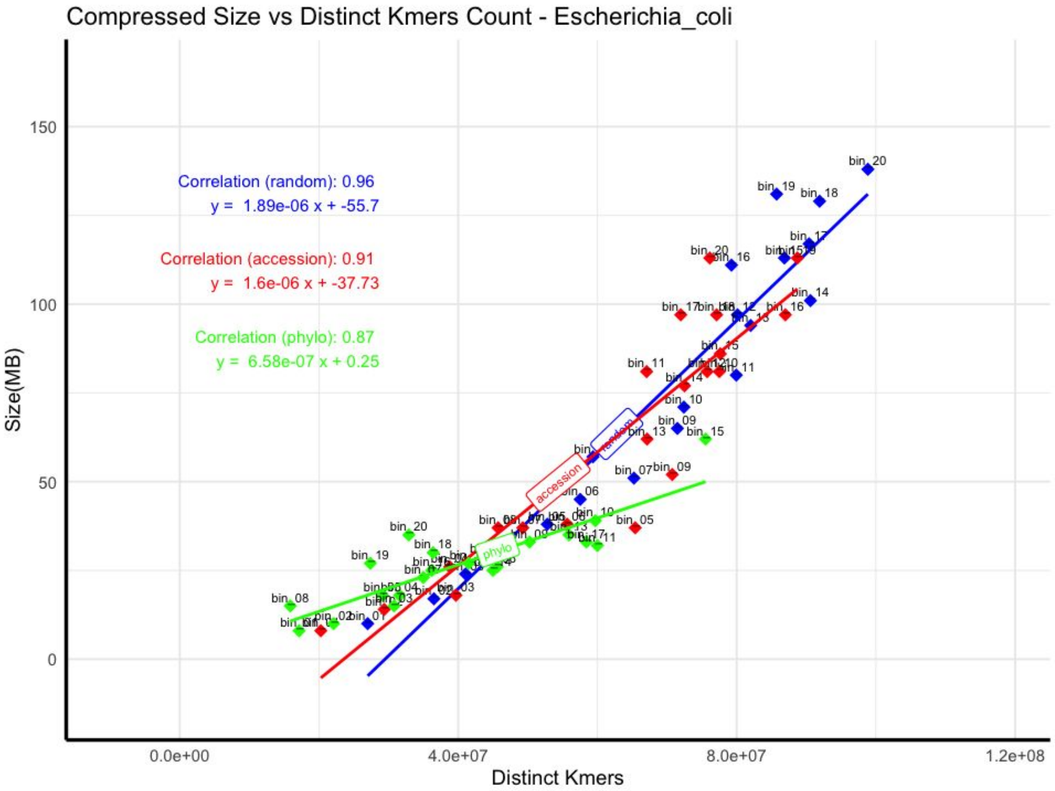
Clostridioides_difficile



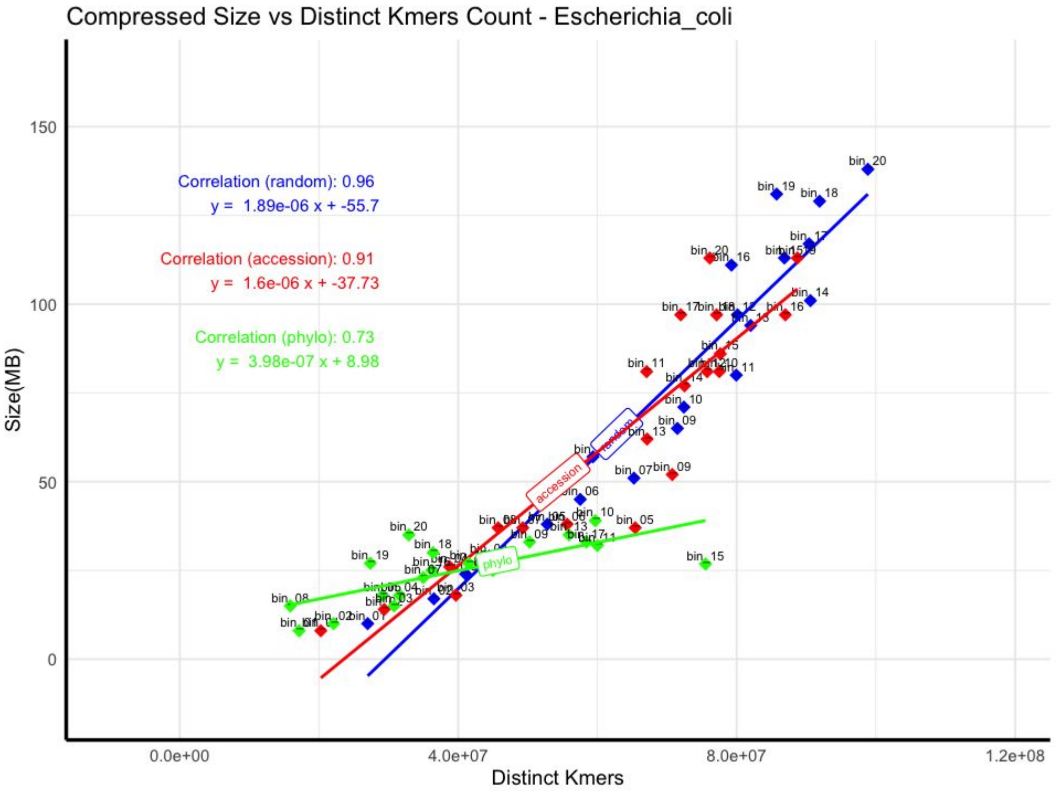
Clostridioides_difficile - replace outliers with means



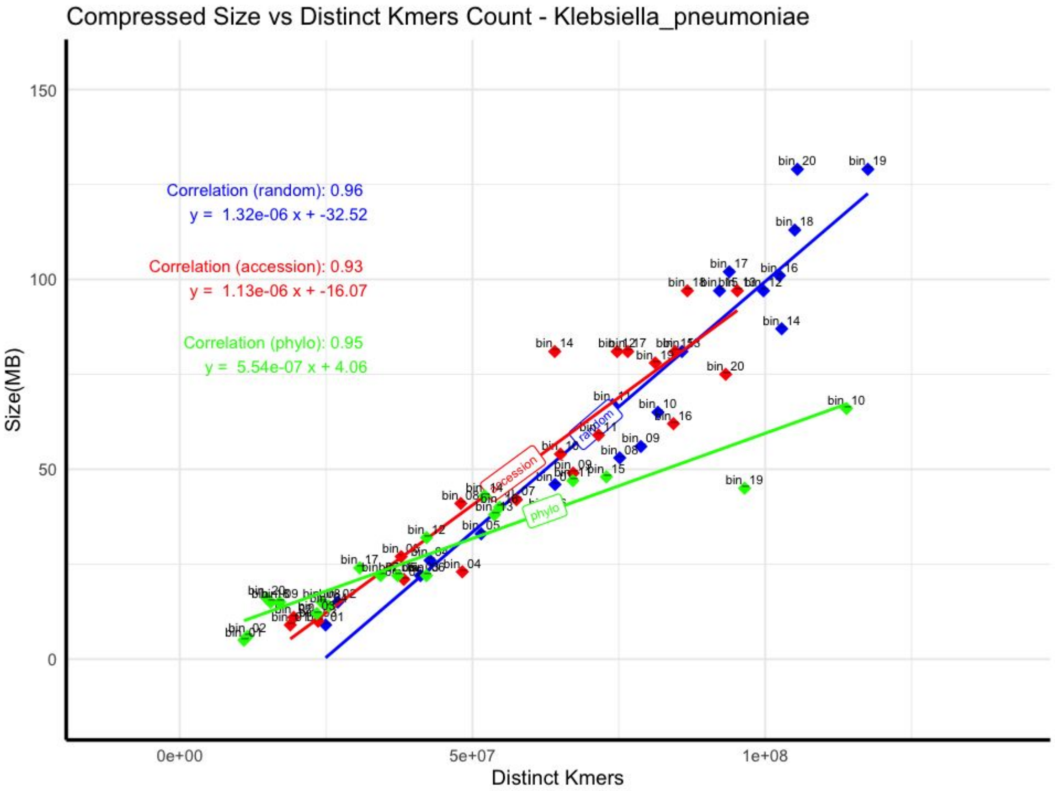
Escherichia_coli



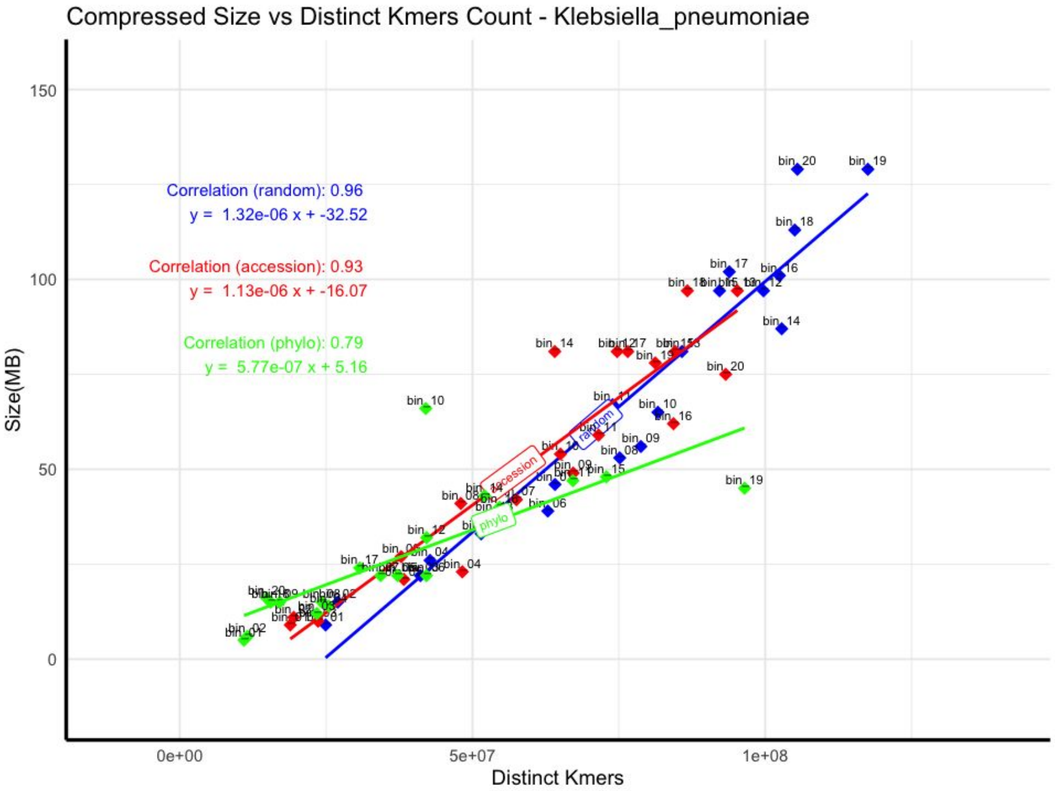
Escherichia_coli - replace outliers with means (no change)



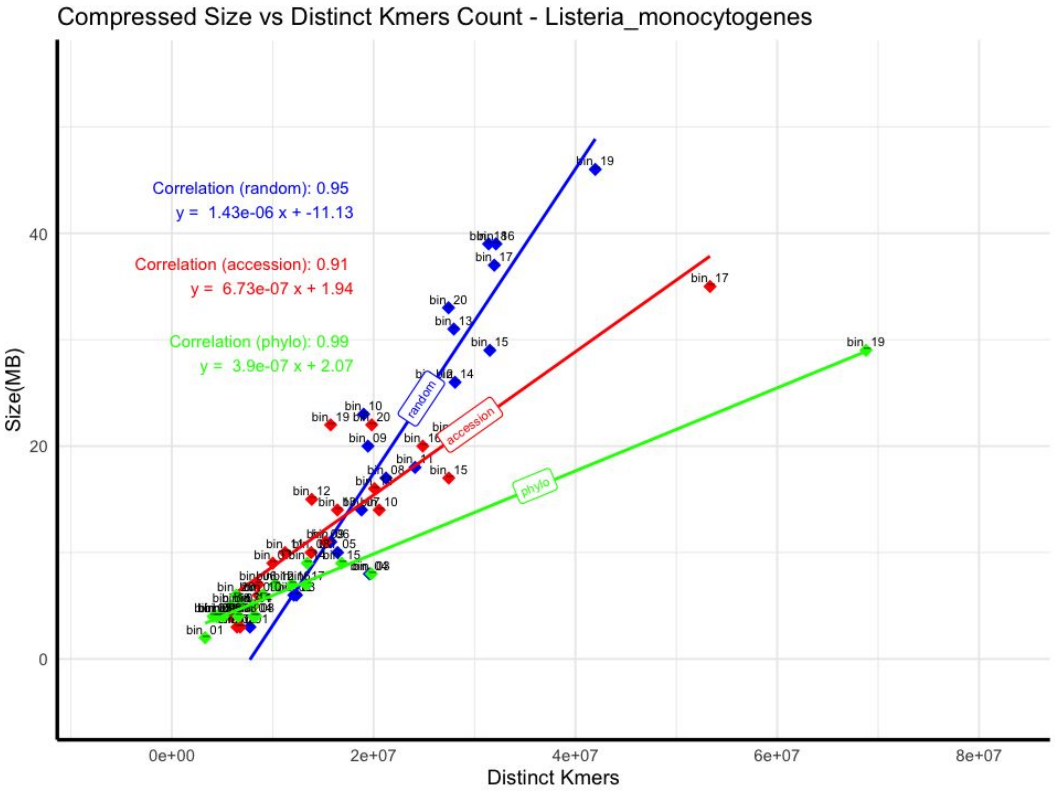
Klebsiella_pneumoniae



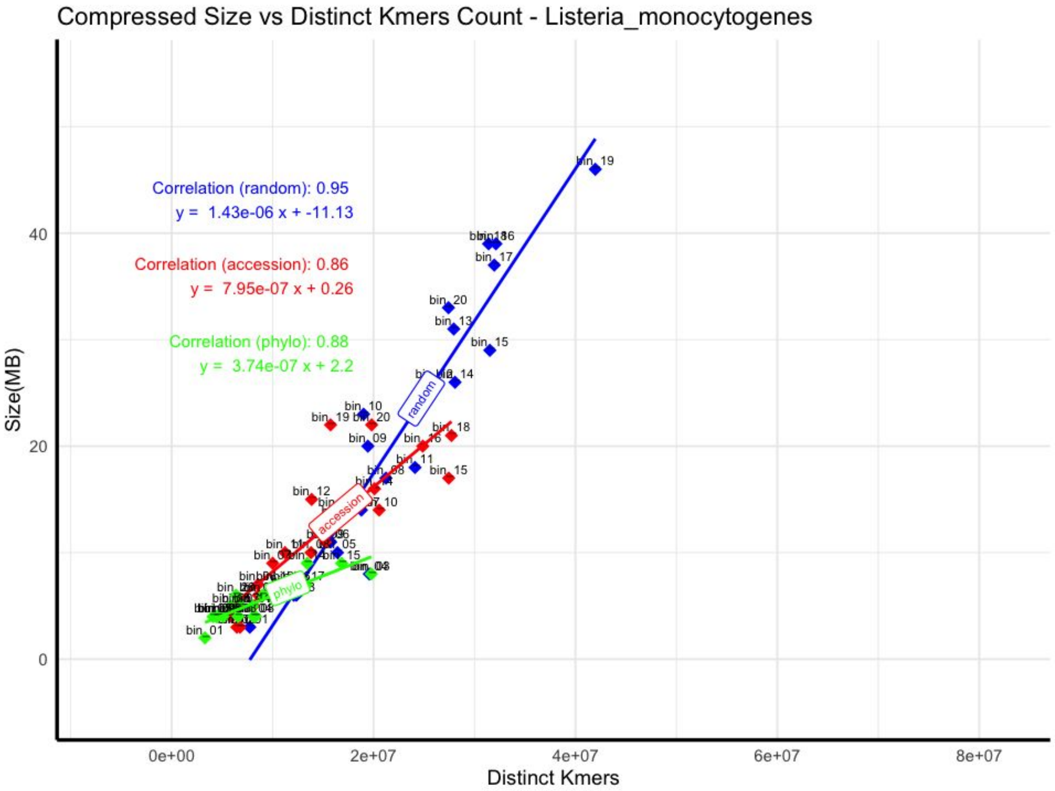
Klebsiella_pneumoniae - replace outliers with means



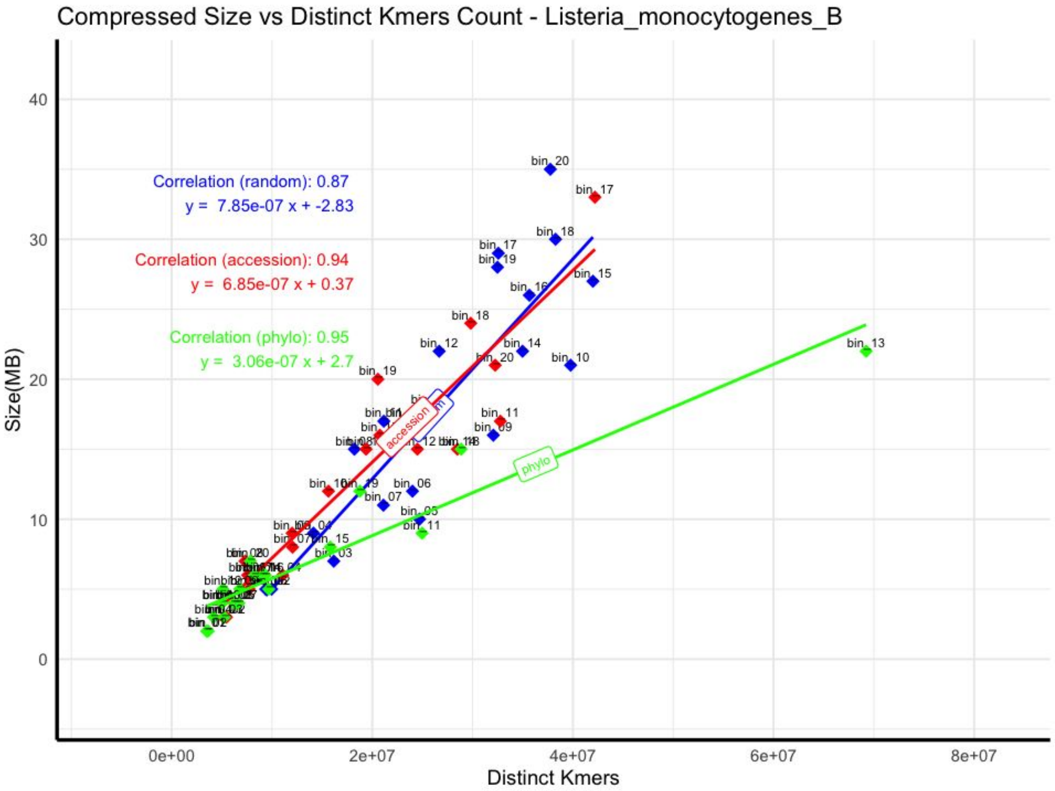
Listeria_monocytogenes



Listeria_monocytogenes - replace outliers with means

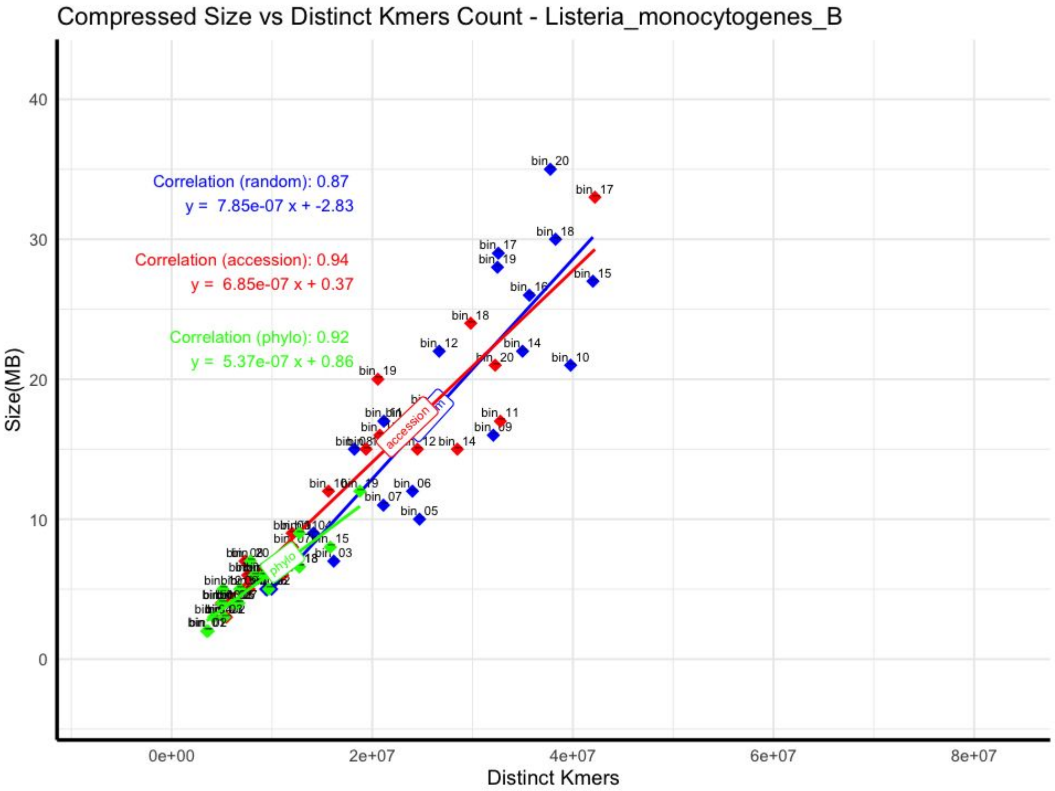


Listeria_monocytogenes_B

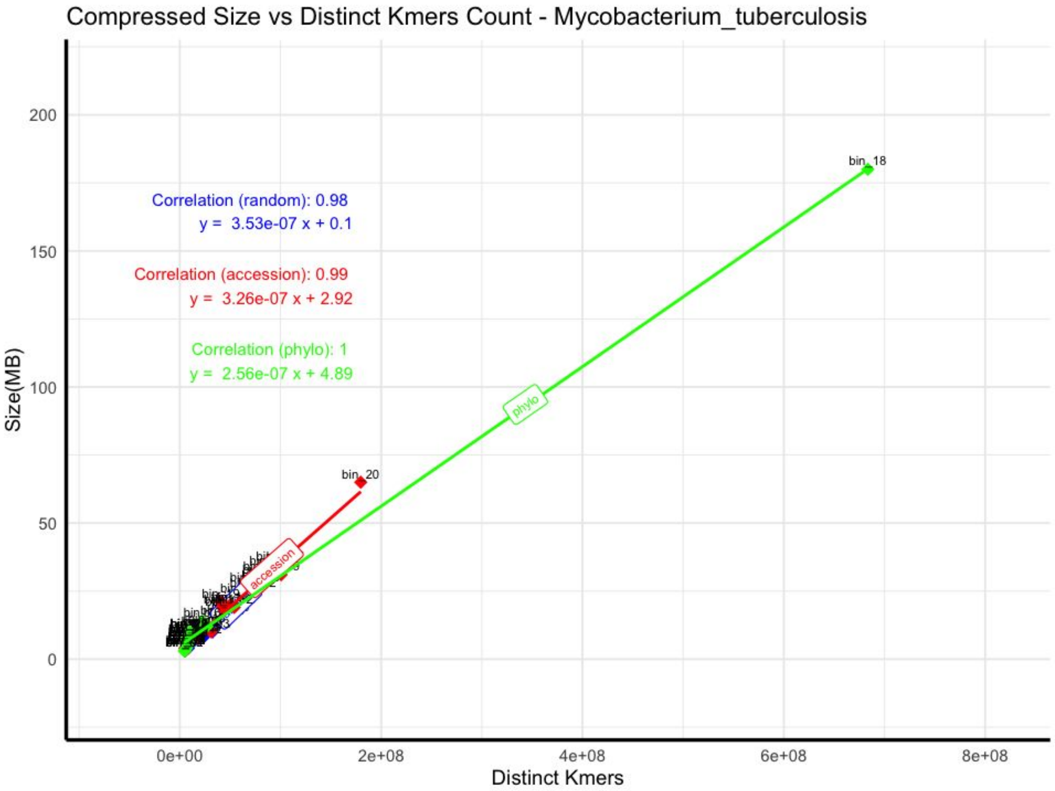


Compressed Size vs Distinct Kmers Count - Listeria_monocytogenes_B

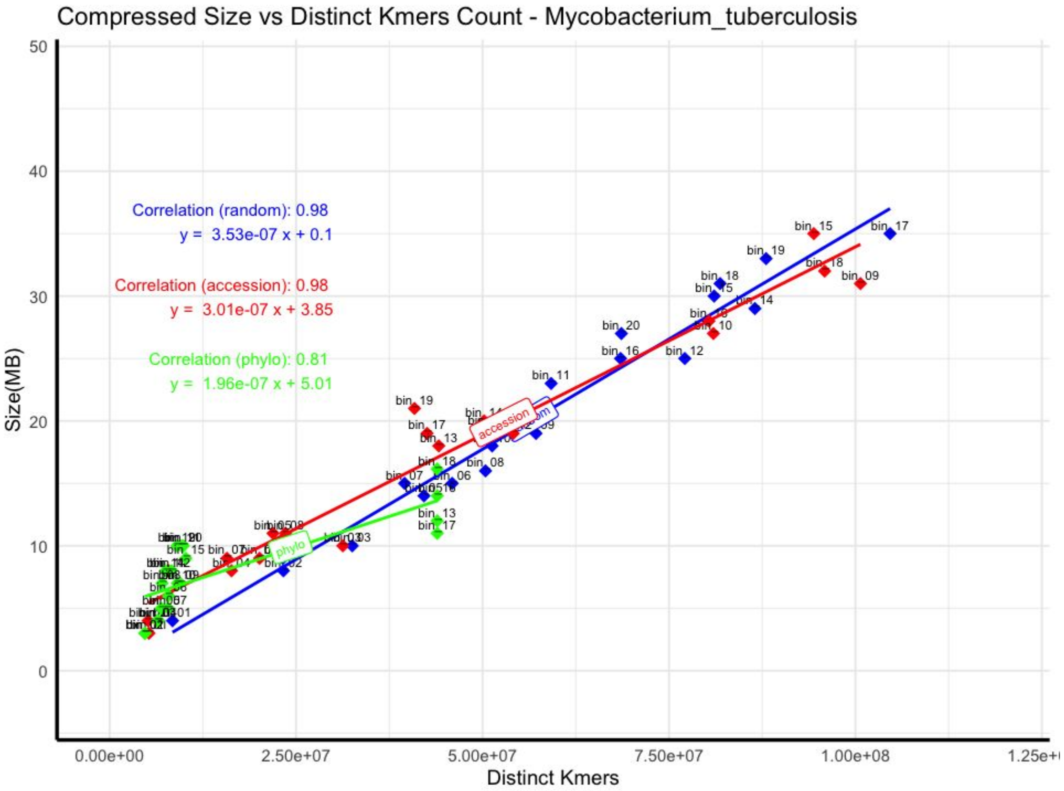
Listeria_monocytogenes_B - replace outliers with means



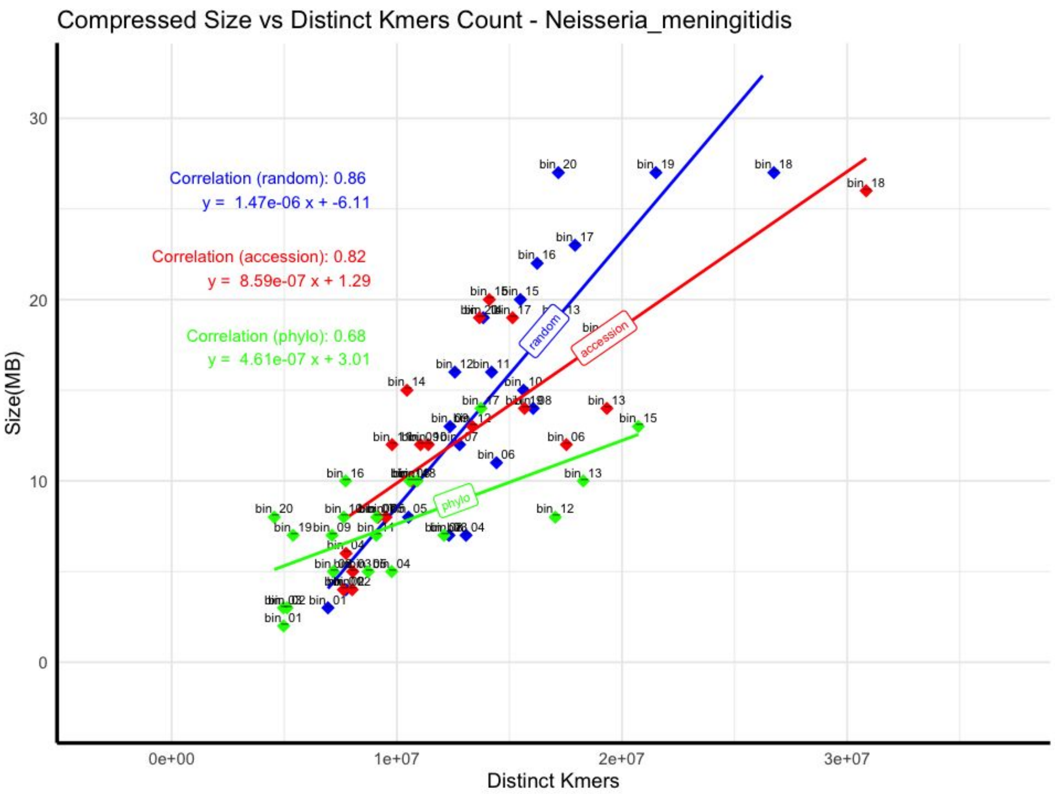
Mycobacterium_tuberculosis



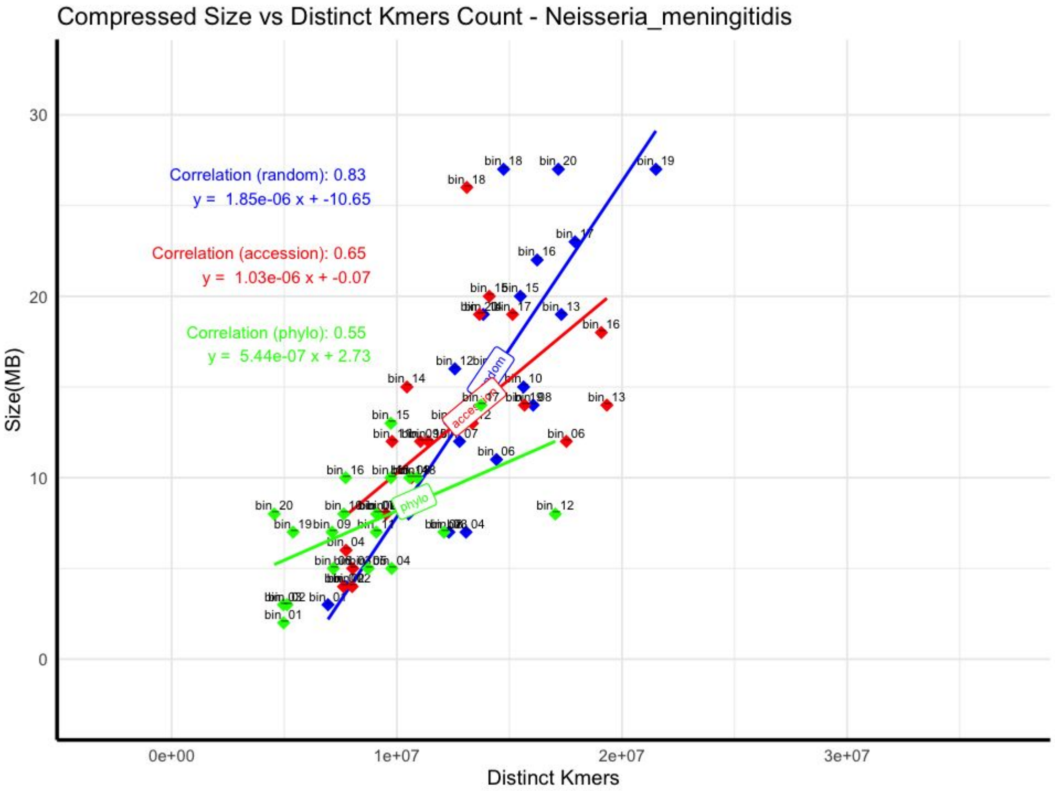
Mycobacterium_tuberculosis - replace outliers with means



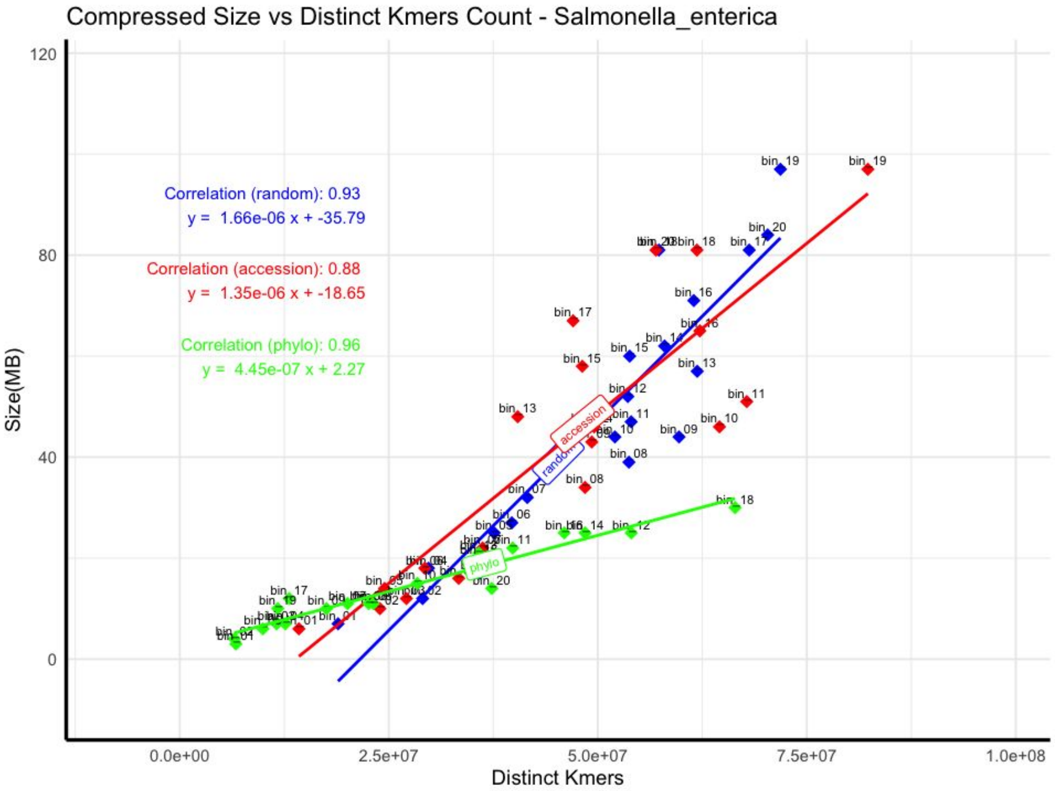
Neisseria_meningitidis



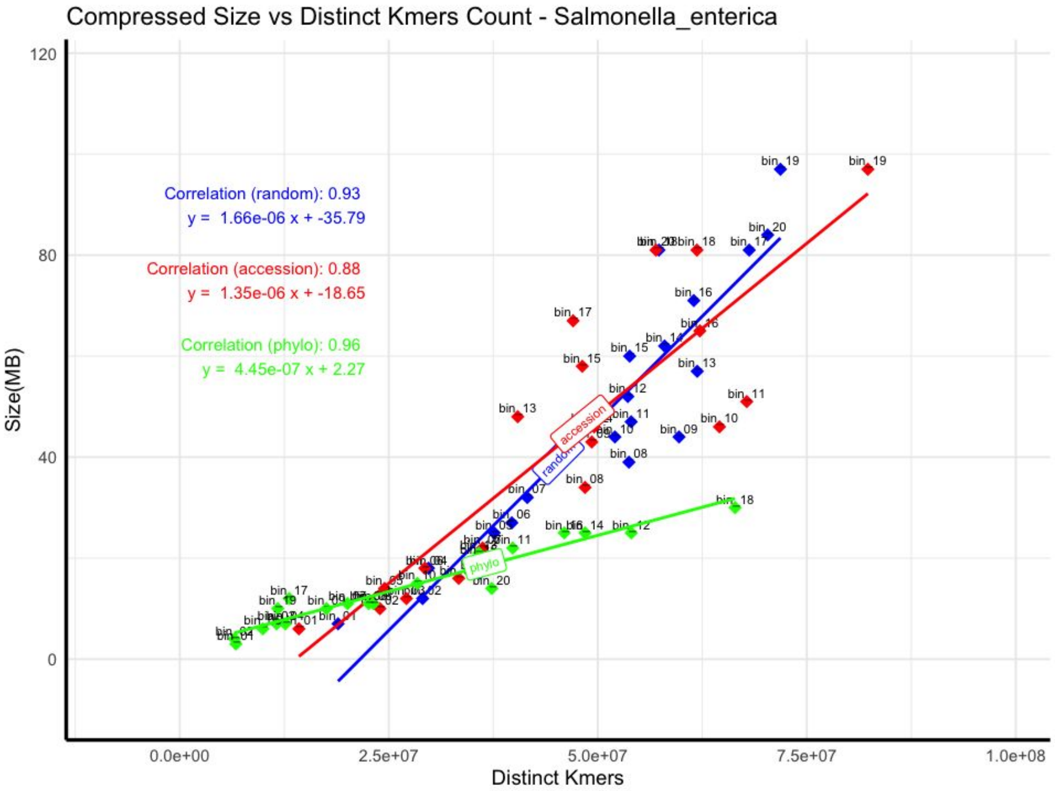
Neisseria_meningitidis - replace outliers with means



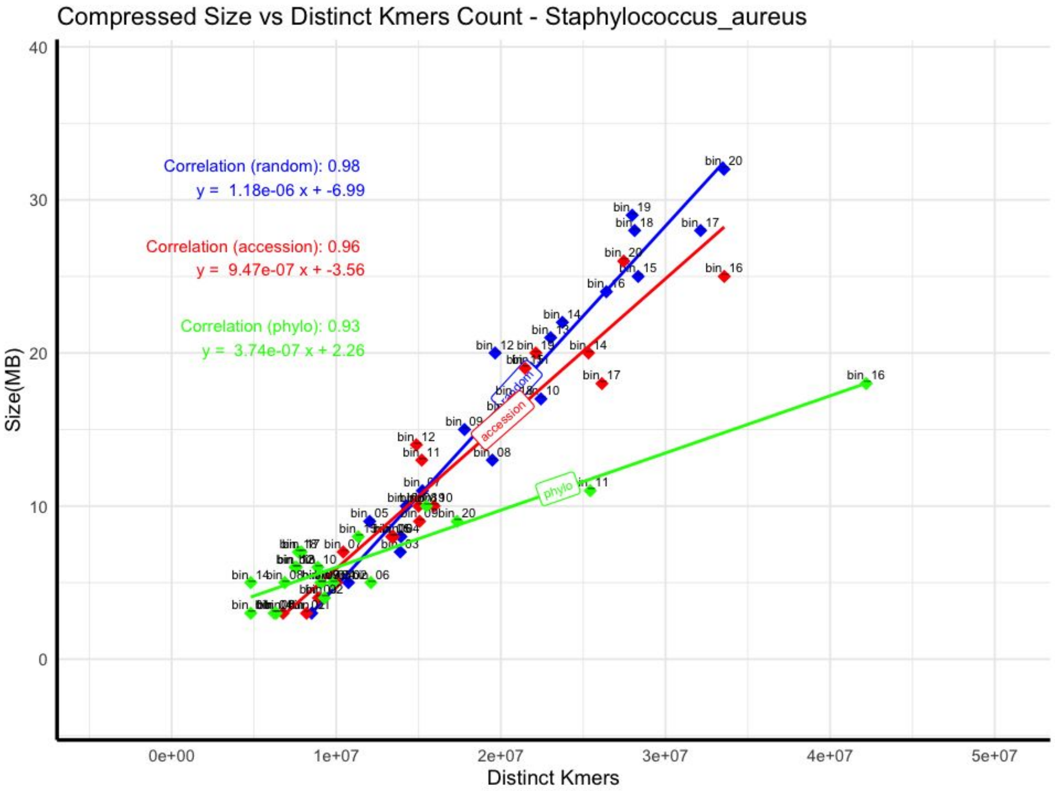
Salmonella_enterica



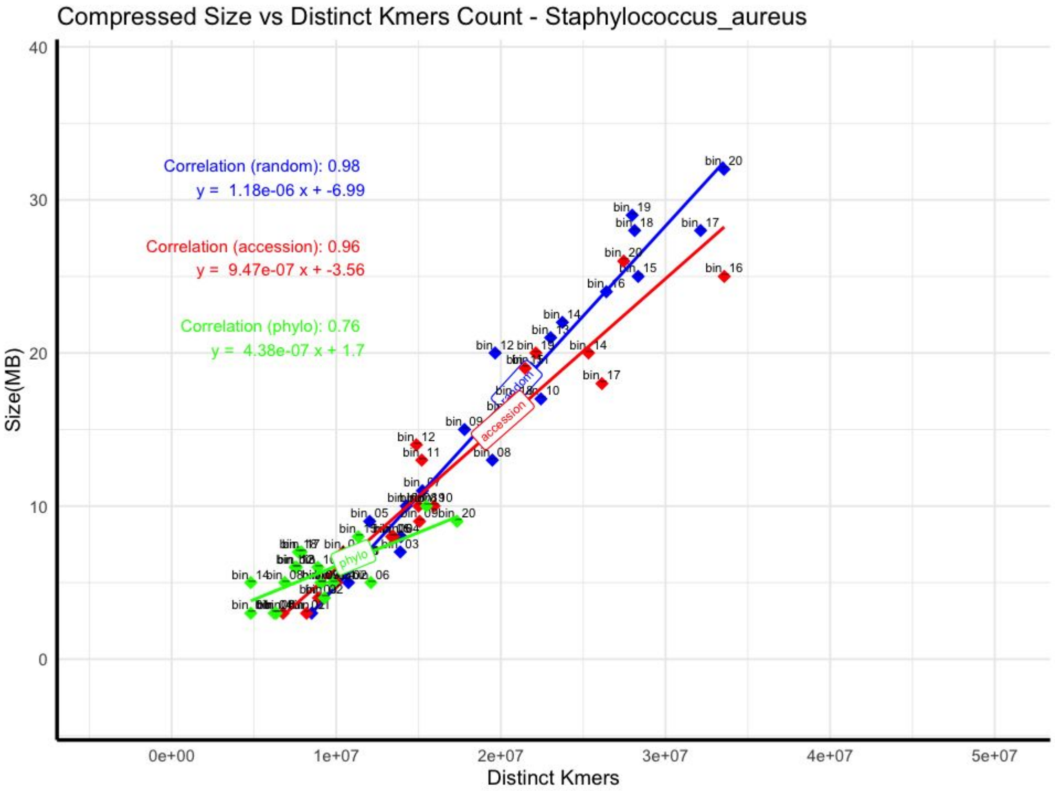
Salmonella_enterica - replace outliers with means (no change)



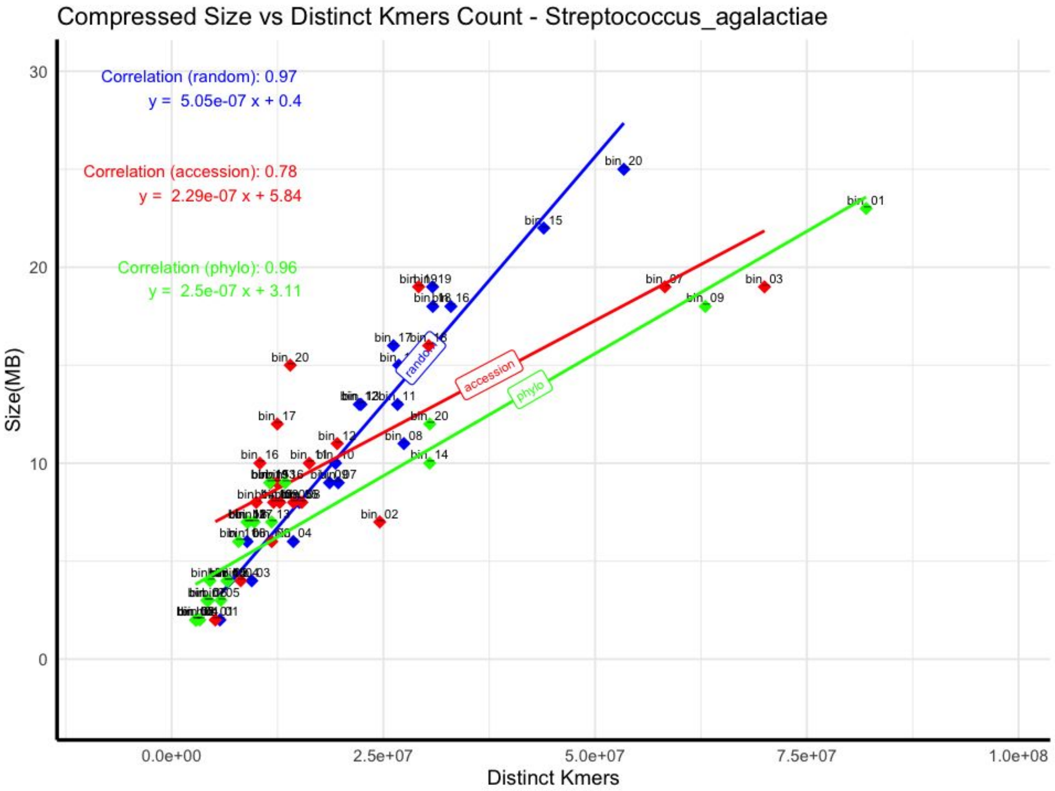
Staphylococcus_aureus



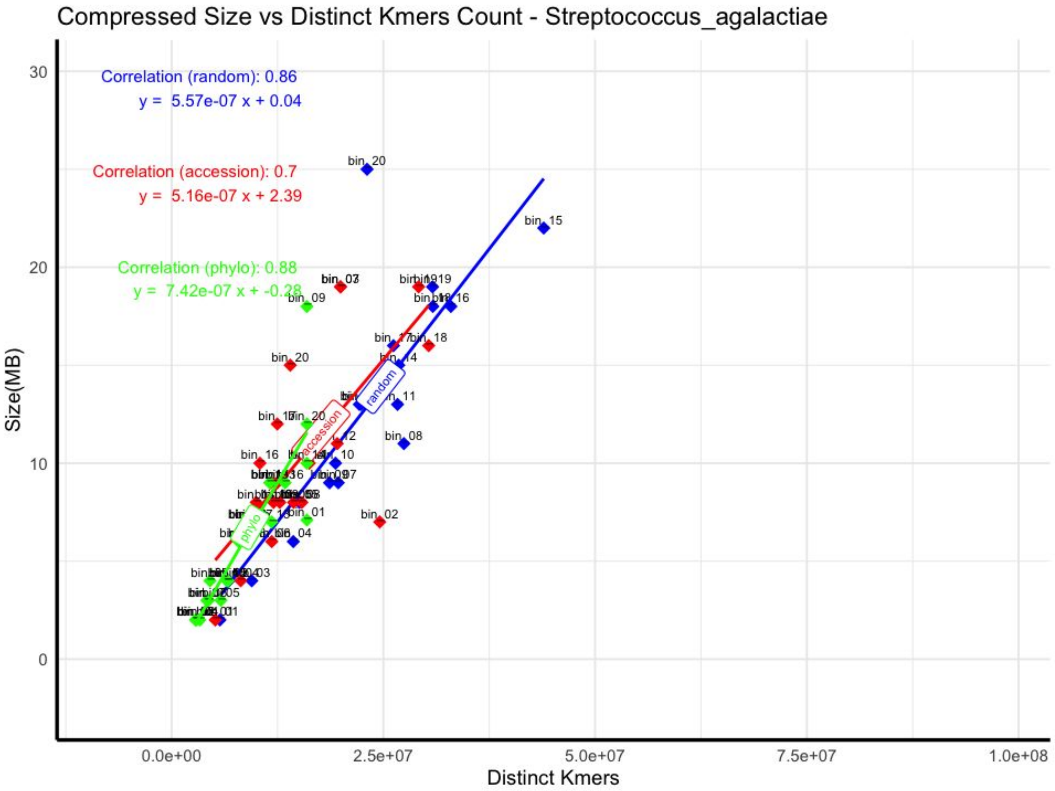
Staphylococcus_aureus - replace outliers with means



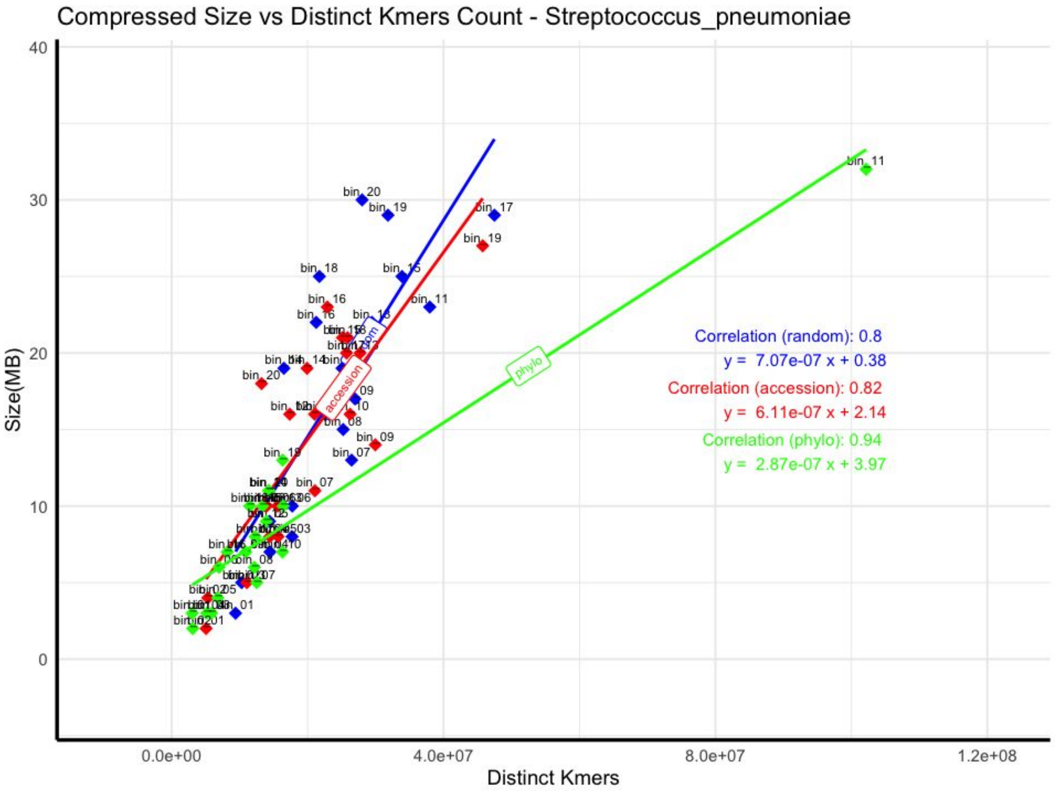
Streptococcus_agalactiae



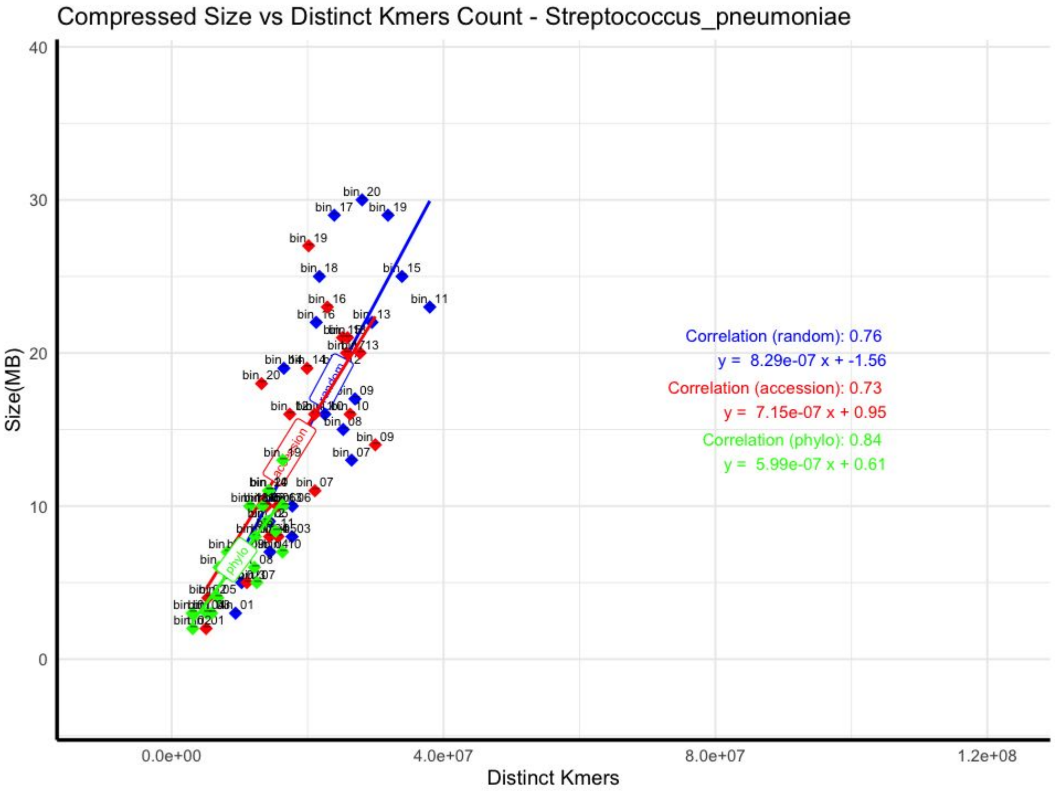
Streptococcus_agalactiae - replace outliers with means



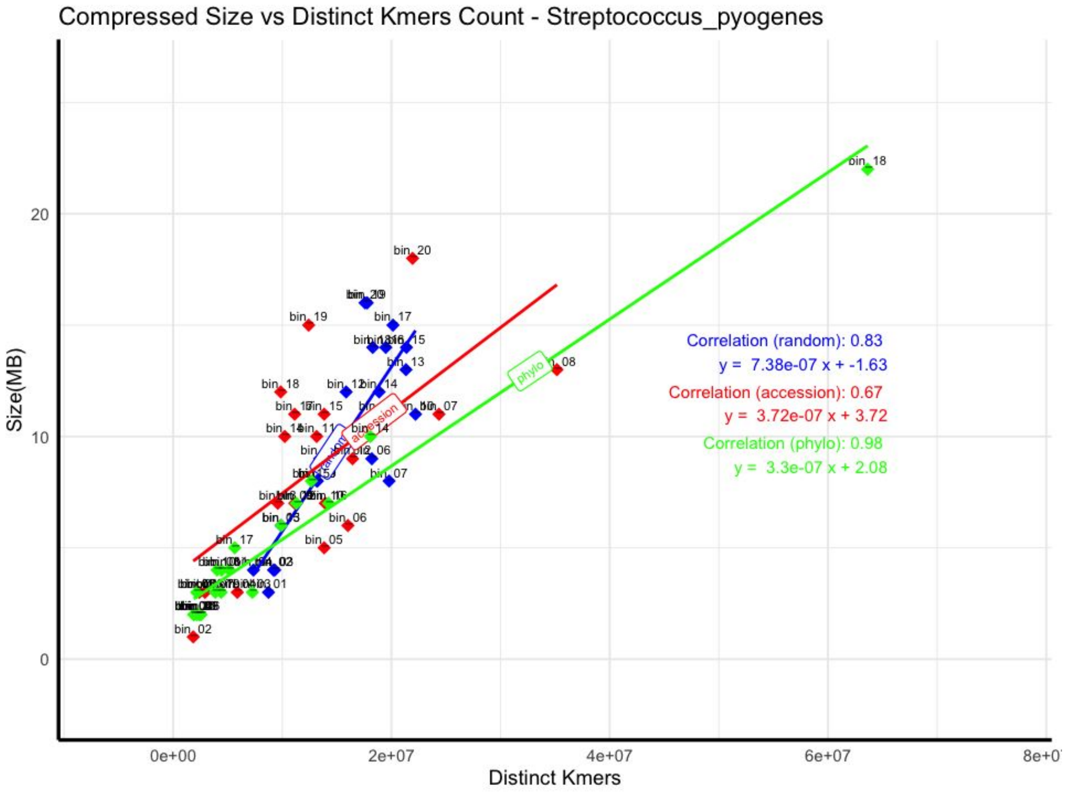
Streptococcus_pneumoniae



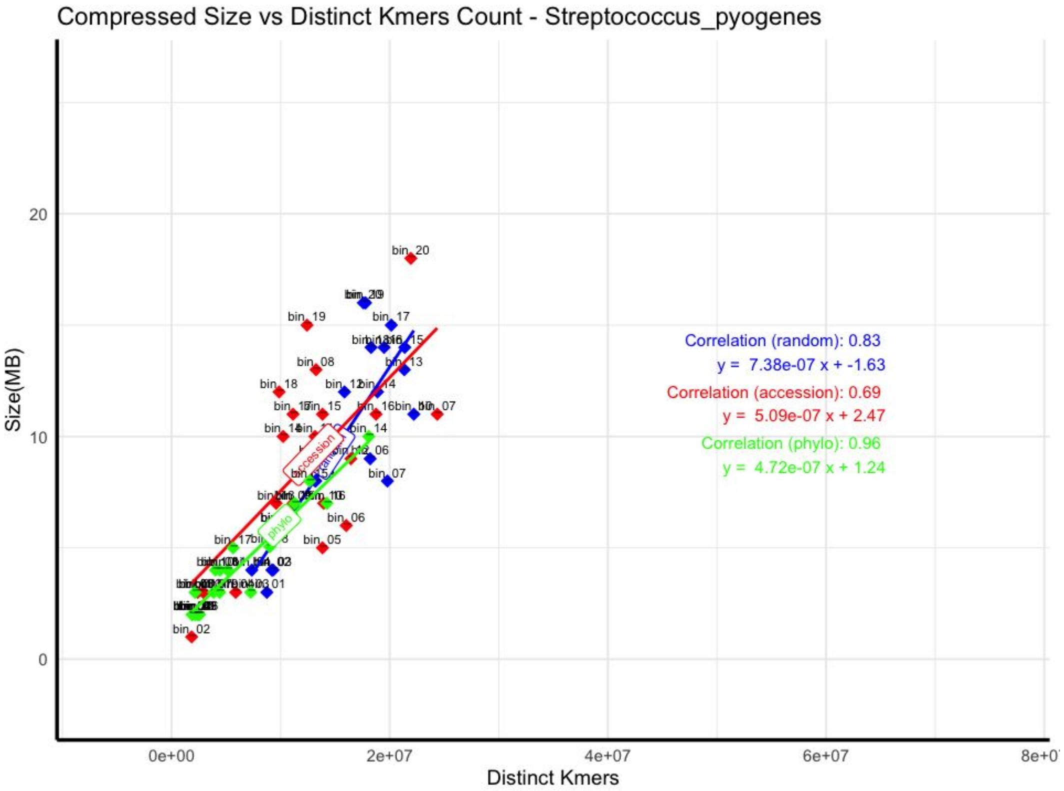
Streptococcus_pneumoniae - replace outliers with means



Streptococcus_pyogenes



Streptococcus_pyogenes - replace outliers with means


Compressed Size vs Distinct Kmers Count - Streptococcus_pyogenes
Size(MB)
Distinct Kmers
0e+00
2e+07
4e+07
6e+07
8e+07
bin_20
bin_19
bin_17
bin_08
bin_13
bin_18
bin_12
bin_14
bin_15
bin_16
bin_07
bin_06
bin_05
bin_02
bin_03
bin_01
accession
random
phylo
Correlation (random): 0.83
y = 7.38e-07 x + -1.63
Correlation (accession): 0.69
y = 5.09e-07 x + 2.47
Correlation (phylo): 0.96
y = 4.72e-07 x + 1.24