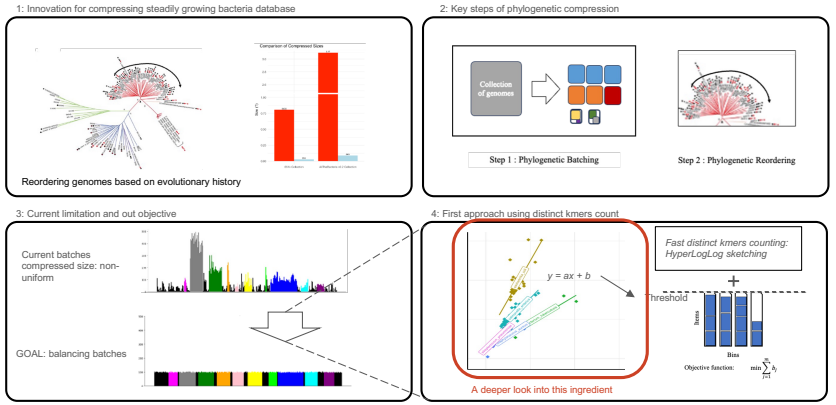
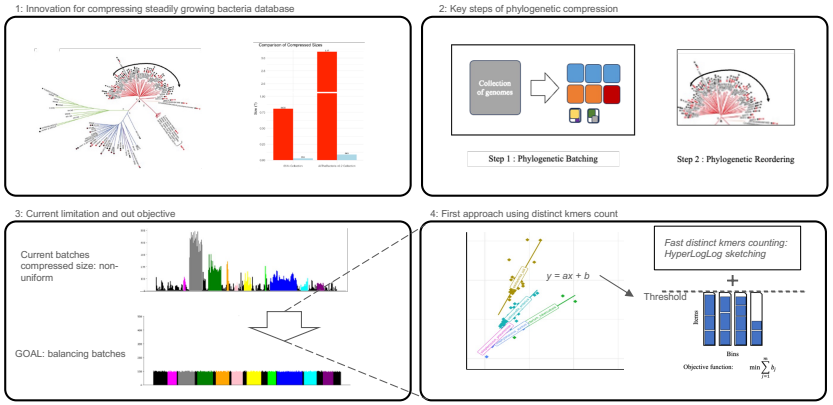
Distinct K-mers Count and Compression Size: Correlations Across Genome Orders

31 Mars 2025

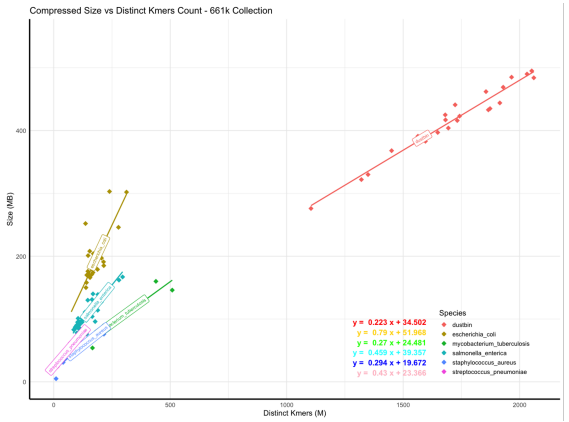
Phylogenetic compression and 3 ingredients for balancing batches



Phylogenetic compression and 3 ingredients for balancing batches



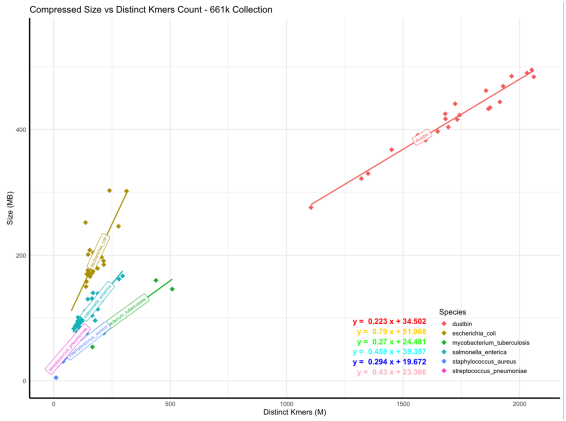
Evaluating post compression sizes prediction using 661k batches



Test data: 10 randomly sampled batches from a single species, each with a varying number of genomes (100 => 1000).

Root Mean Squared Error is used in this presentation for ease of interpretation

Evaluating post compression sizes prediction using 661k batches



Test data: 10 randomly sampled batches from a single species, each with a varying number of genomes (100 => 1000).

Salmonella enterica:
Root Mean Squared Error (RMSE): 34.850 MB

Escherichia coli:
Root Mean Squared Error (RMSE): 53.911 MB

Mycobacterium tuberculosis:
Root Mean Squared Error (RMSE): 26.651 MB

Not particularly good prediction

⇒ Possible reason: The batches in the 661k dataset are not random; they have a specific ordering.

Does accounting for genome orderings improve the approximation of post-compression sizes?

Experiment setup - Dataset:

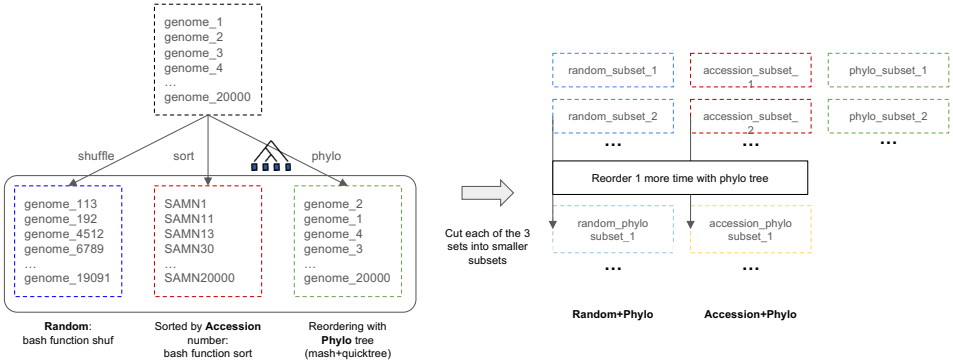
From the 661k genome collection, select **6 species** (gtdbtk classification) **with over 20,000 genomes.**

Sample randomly 20,000 genomes from each species.

The 6 selected species comprise **444,507** genomes, or **67% of the collection.** We sampled **120,000** genomes (20,000 per species), representing **18% of the total collection.**

Species: *Campylobacter_D jejuni, Escherichia coli, Mycobacterium tuberculosis , Salmonella enterica, Staphylococcus aureus, Streptococcus pneumoniae*

Orders: Random, Accession, Phylogenetic, Accession+Phylo, Random+Phylo



Split the genomes into different size groups

The set genomes are then cut into subsets, starting with a small number of genomes and gradually increasing the number.

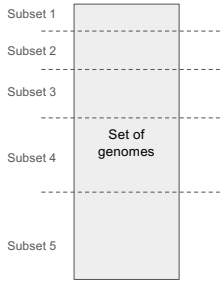
Example: with 20 genomes, and 4 groups:

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

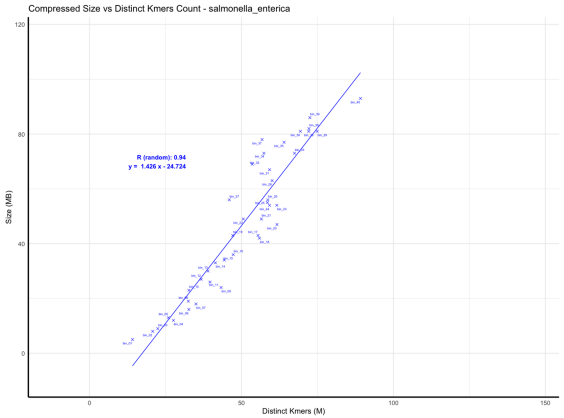
We cut the set of **20,000 genomes of single species into 40 disjunctive subsets.**

The smallest one has **24**, the largest has **976** genomes

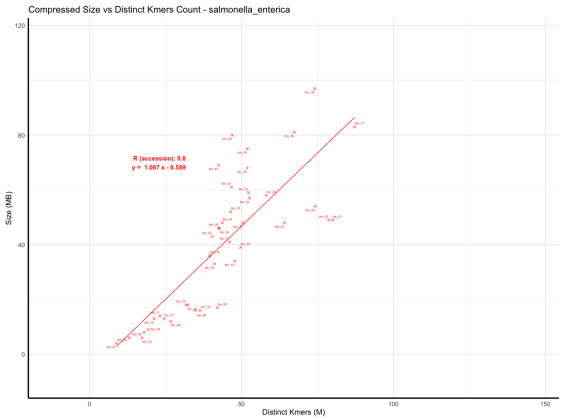
Then we estimate the distinct kmers and compress all of them.



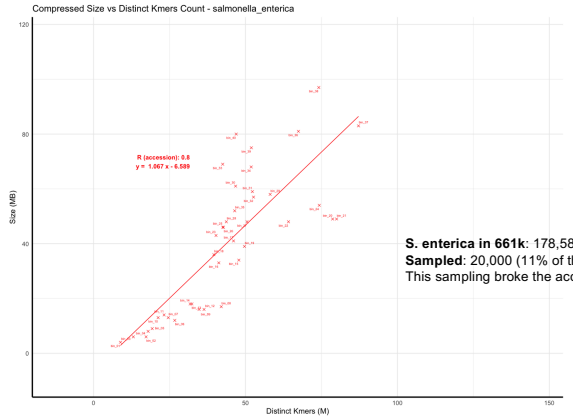
Linear regression result: Salmonella enterica - random order



Linear regression result: Salmonella enterica - accession order

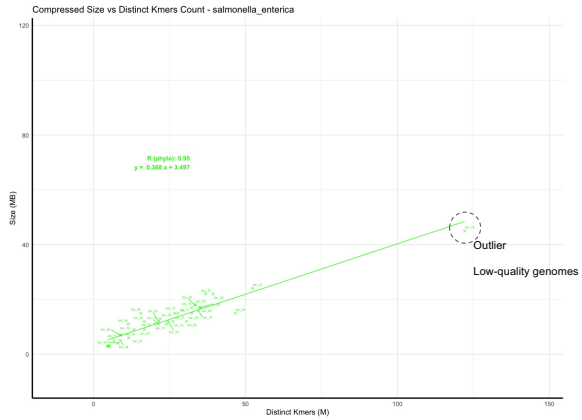


Linear regression result: Salmonella enterica - accession order

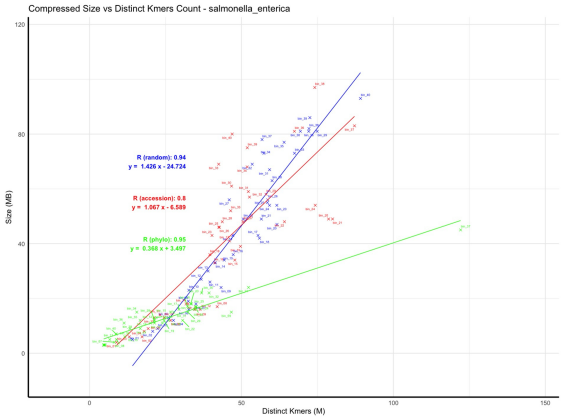


S. enterica in 661k: 178,585 genomes
Sampled: 20,000 (11% of the total)
This sampling broke the accession order.

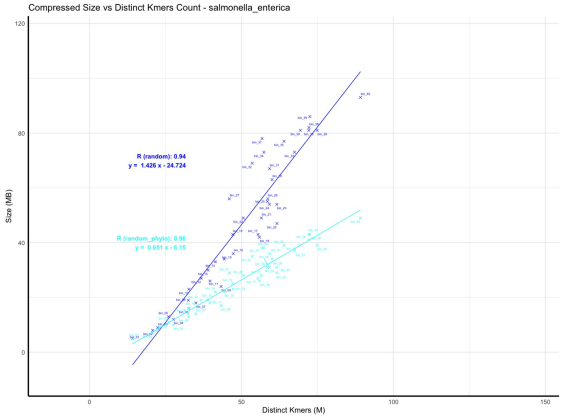
Linear regression result: Salmonella enterica - phylo order



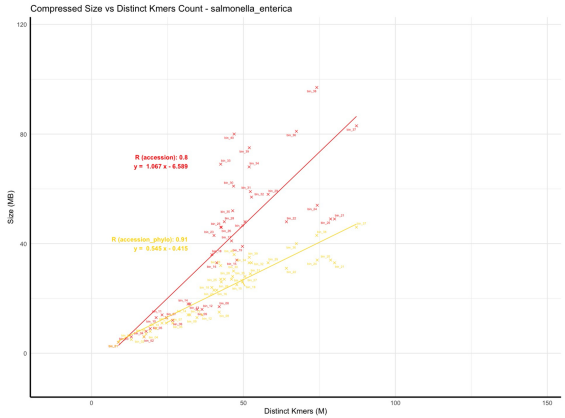
Linear regression result: Salmonella enterica - random, accession, phylo



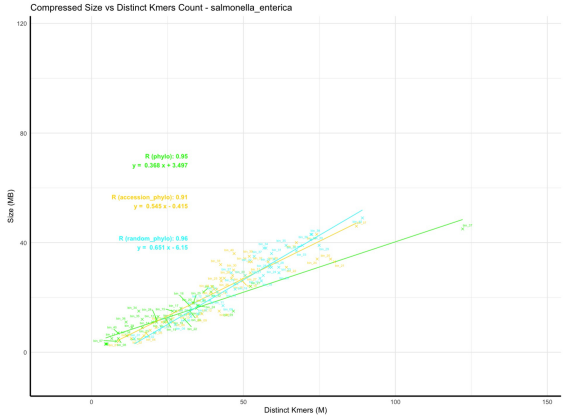
Linear regression result: Salmonella enterica - random, random + phylo



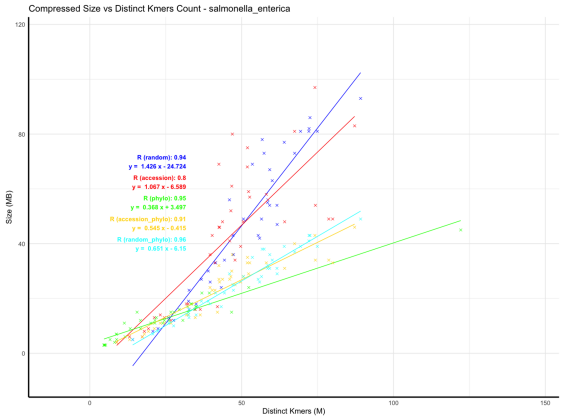
Linear regression result: Salmonella enterica - accession, accession + phylo



Linear regression result: Salmonella enterica - phylo, accession+phylo, random+phylo



Linear regression result: Salmonella enterica - all orders



	RMSE (MB)
Random	45.623
Accession	36.236
Phylo	17.533
A+Phylo	14.345
R+Phylo	17.205

Summary: RMSE (MB) of 6 species investigated, how well can we predict?

	S.enterica	E.coli	M.tuberculosis	S.aureus	C.D.jejuni	S.pneumoniae
Random	45.623	53.375	20.590	Still compressing		
Accession	36.236	43.26	26.810			
Phylo	17.533	32.325	39.635			
A+Phylo	14.345	21.464	17.702			
R+Phylo	17.205	23.332	13.761			

Discussion and perspectives

Improved Approximation (3 of the most highly sampled species for now): Considering species and genome ordering improves post-compression size approximation.

Phylogenetic Compression: Its impact grows as batch size increases.

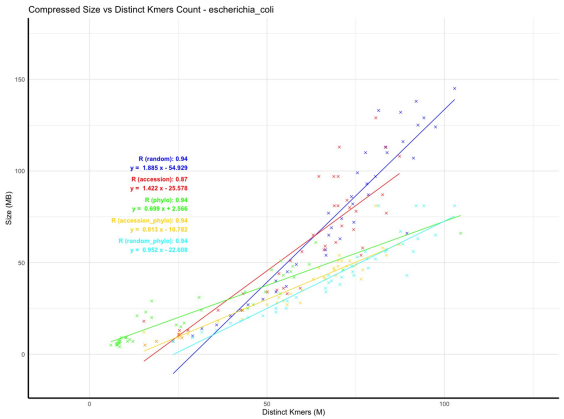
Scalability: Estimating a phylogenetic tree for 20,000 genomes is feasible—can we scale to 100,000 genomes or more?

Challenge: Outliers negatively impact results.

Future Improvement: Incorporate genome quality to handle outliers effectively.

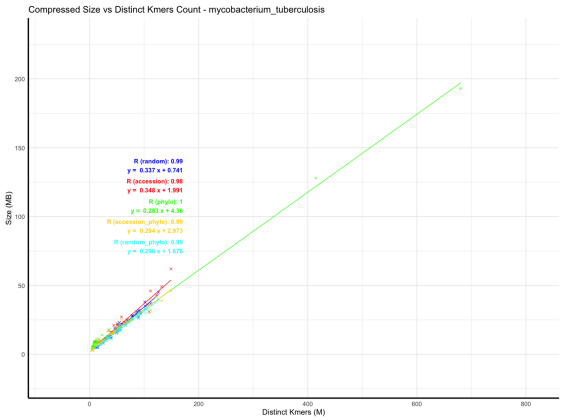
Supplementary

escherichia_coli



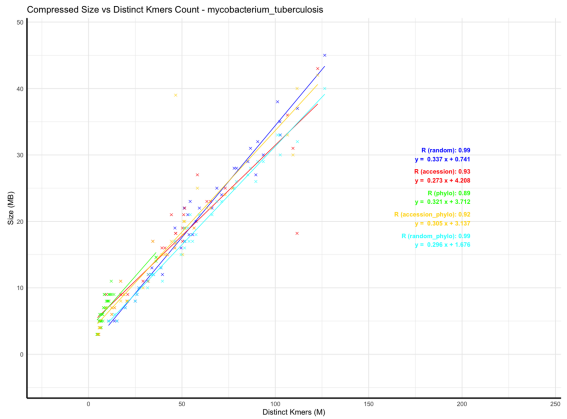
21

mycobacterium_tuberculosis



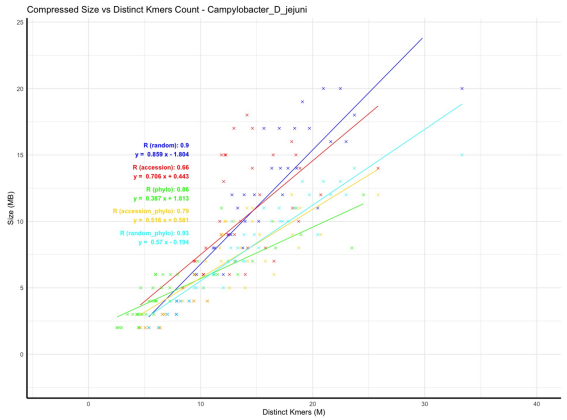
22

mycobacterium_tuberculosis; replaced outliers with means



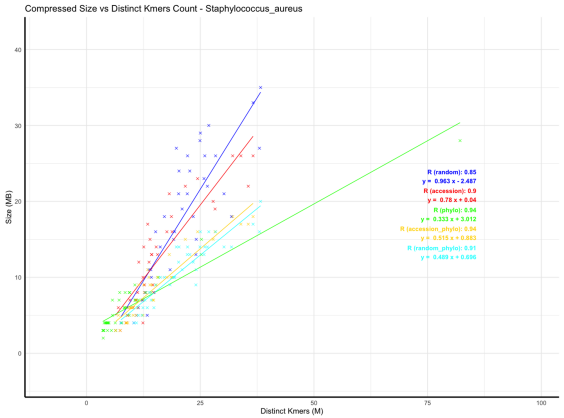
23

Campylobacter_D_jejuni

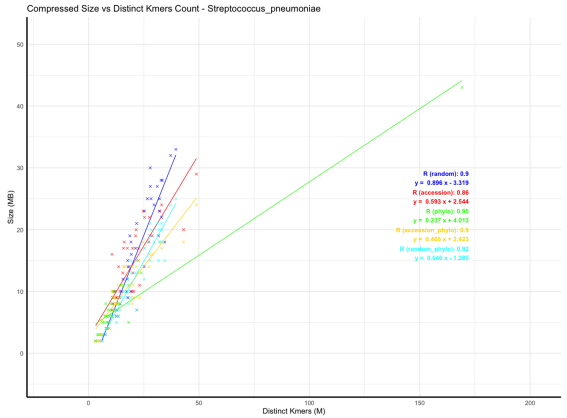


24

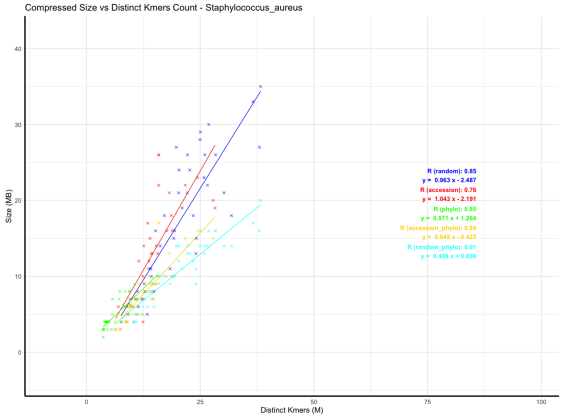
Staphylococcus_aureus



Streptococcus_pneumoniae



Staphylococcus_aureus; replaced outliers with means



Streptococcus_pneumoniae; replaced outliers with means