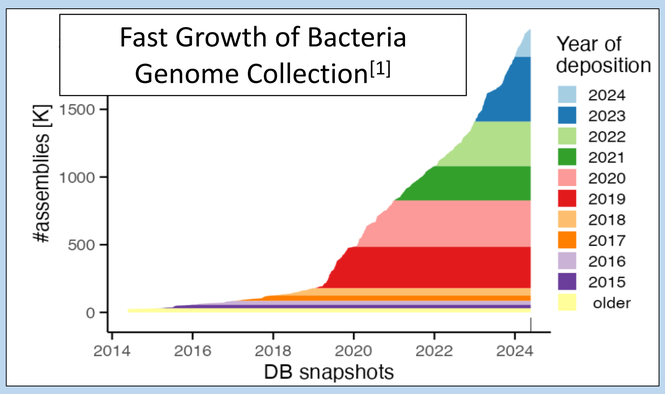
HyperLogLog-Based Load Balancing and Bin Packing for Efficient Compression of Large Bacterial Genomes Collection

Tam TRUONG, Dominique LAVENIER, Pierre PETERLONGO, Karel BRINDA

IRISA, Inria, GENSCALE, Université de Rennes

MOTIVATION



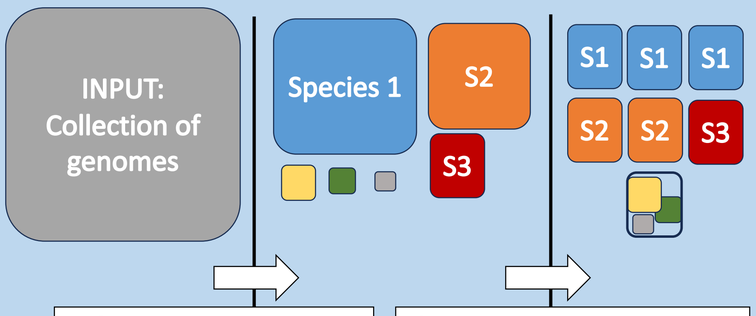
Large Bacterial Genome Collections:

661k collection[2] (2021)	n = 661,405
AllTheBacteria[3] (2024)	n = 2,440,377
Future	n > 10^7

Goal: efficient compression and search

RECENT INNOVATION: PHYLOGENETIC COMPRESSION

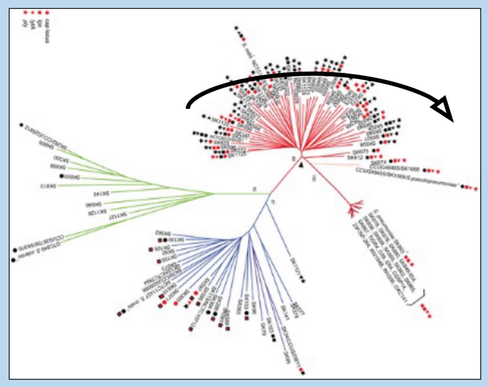
STEP 1 : PHYLOGENETIC BATCHING



Species clustering

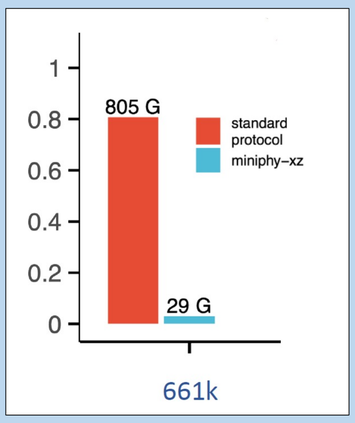
Split or merge based on cluster size (batching)

STEP 2: PHYLOGENETIC REORDERING



Reordering of genomes in each batch using an estimated evolutionary tree

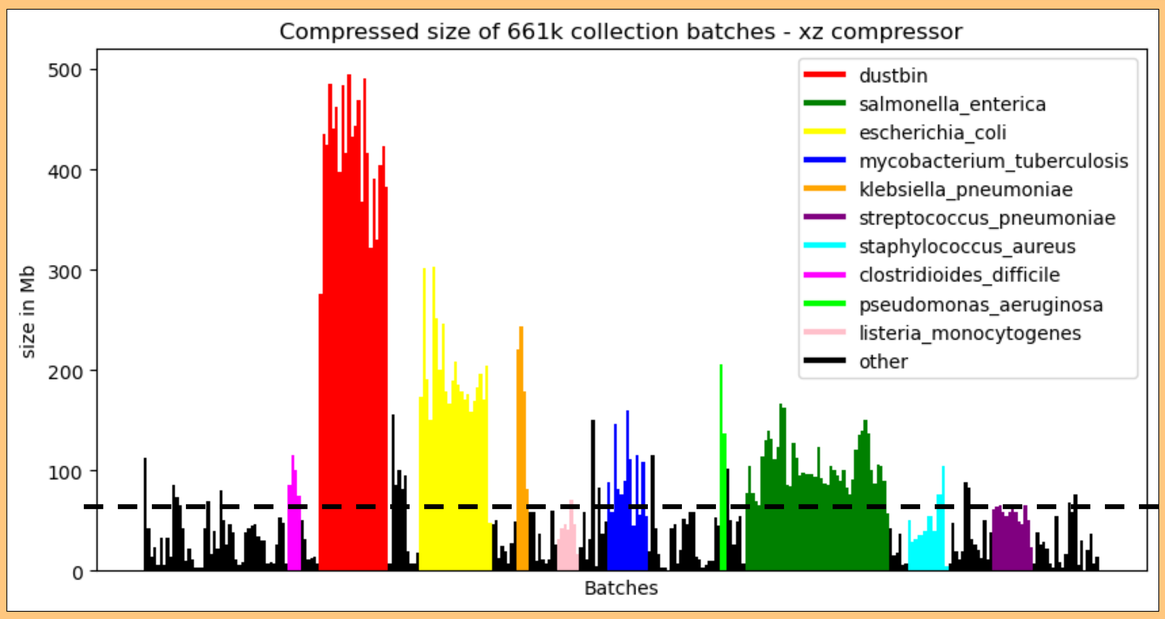
RESULTING COMPRESSION



Lossless compression of 1-3 orders of magnitude

Key idea: improves compressibility via reordering according to the evolutionary history

CURRENT LIMITATION: Batching Results In Non-uniform Compressed Sizes



CONSEQUENCES

- Unbalanced Workloads
- Inefficient Data Transmission
- Hinder Parallelization
- Inconsistent Query Times
- Memory Overuse

ULTIMATE OBJECTIVE

Objective :

$$min \sum resource(batch)$$

Per-batch Constraints :

Compressed size
Decompressed size
Number of genomes
Search indexes size

Applications:

Portable Devices
(Remote setting, field work, rapid diagnostic)

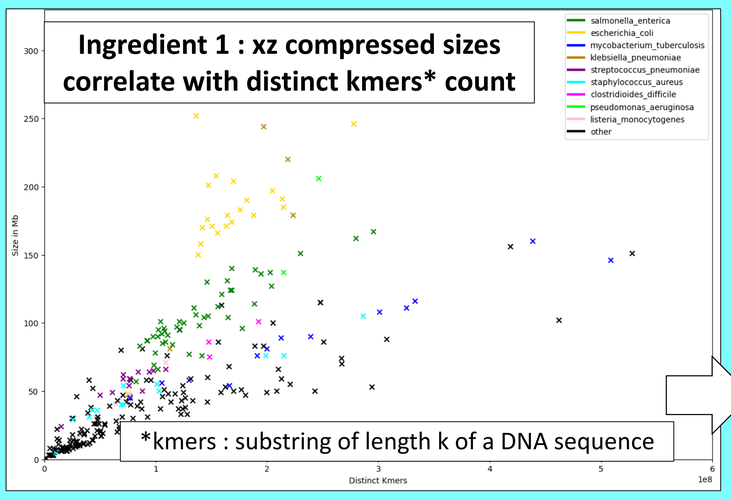
Parallel Platforms
(GPU, Processing-in-Memory)

CURRENT GOAL

Balance post-compression batches for rapid and reliable internet transmission (threshold on post-compression batch size)

METHODS

Ingredient 1 : xz compressed sizes correlate with distinct kmers* count



*kmers : substring of length k of a DNA sequence

Ingredient 2: Cardinality estimation using HyperLogLog sketching

Sketches : approximate data structures.

HyperLogLog sketches for cardinality est.: bit patterns,

i.e. *hash(ATGCG)* ➔ 00010100, *hash(CGTAC)* ➔ 00000010.

Fast and efficient UNION operation for sketches.

Prediction of Genome Batch Post-Compression Size Via Distinct Kmers Estimation

Ingredient 3: Load Balancing[6] and Bin Packing[7]

Preliminary : Given m genomes, put genomes into batches :

STRATEGY 1 : given unlimited batches with capacity C

$$\boldsymbol{Minimize\ nb\ of\ batch\ B}$$

s.t. $distinct\_kmers(bj) < C,\ \text{for (j} = 1, \dots, \text{n)}$

STRATEGY 2 : given a fixed number of batch n

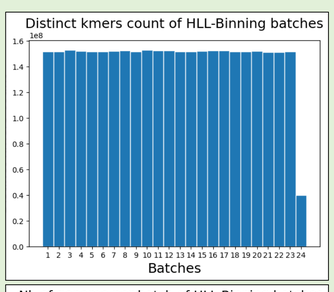
$$\boldsymbol{Minimize\ max\left(distinct\_kmers(b_j)\right), for\ j = 1, \dots, n}$$

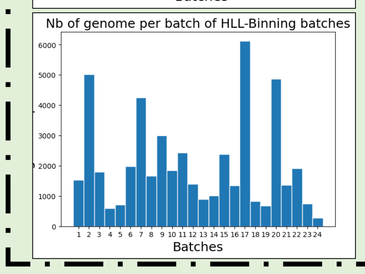
PRELIMINARY RESULTS

DATA : Genomes of *Mycobacterium tuberculosis* from the 661k Collection[2] , B = 24

STRATEGY 1: HLL-Binning

Batches Obtained From Strat. 1

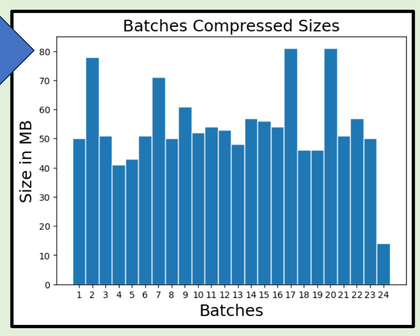




PHYLOGENETIC COMPRESSION

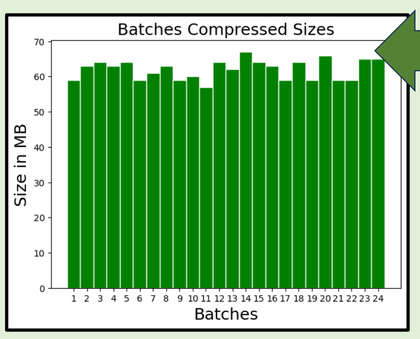
Batch capacity :
C = 152,000,000
(C obtained by linear regression)

Number of genome per batch varies



Most of the batches are balanced (between 40-50MB, max size 81MB)

Evaluation strat. 1:
Allowing a capacity on distinct kmers.
The result remains somewhat imbalanced.



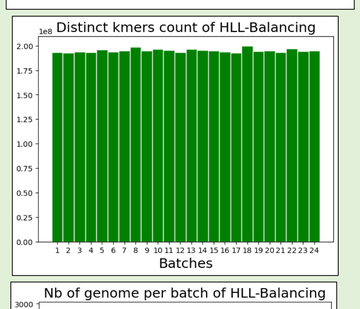
All Batches are well balanced (between 59-67MB, max size 67MB)

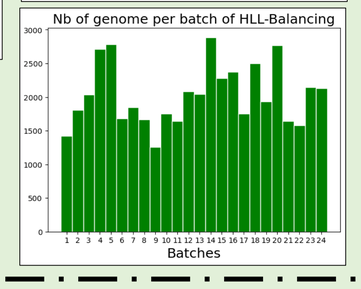
Evaluation strat. 2:
Producing more balanced batches.
No control over the maximum distinct k-mer count per batch.

PHYLOGENETIC COMPRESSION

STRATEGY 2: HLL-Balancing

Batches Obtained From Strat. 2





Nb of genomes per batch varies but to a lesser extent compared to Strat. 1

CONCLUSION & PERSPECTIVES

Batching by Predicting Compression Size Using HyperLogLog Distinct K-mer Estimation Improves balancing of the final compressed sizes *Mycobacterium tuberculosis.*

Current Goals:

- Extending the results and methods to the whole 661k collection.
- Enabling control over the number of genomes in each batch.
- Scaling up to AllTheBacteria collection.
- Applications in querying data structures such as Bloom filter, on PIM and GPU.

BIBLIOGRAPHY

[1] Břinda et al., Efficient and Robust Search of Microbial Genomes via Phylogenetic Compression. To be appeared in *Nature Methods*. 2025

[2] Blackwell et al., Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences. *PLOS Biology* 19, 11. 2021

[3] Hunt et a.l,. AllTheBacteria - all bacterial genomes assembled, available and searchable. *bioRxiv*. 2024

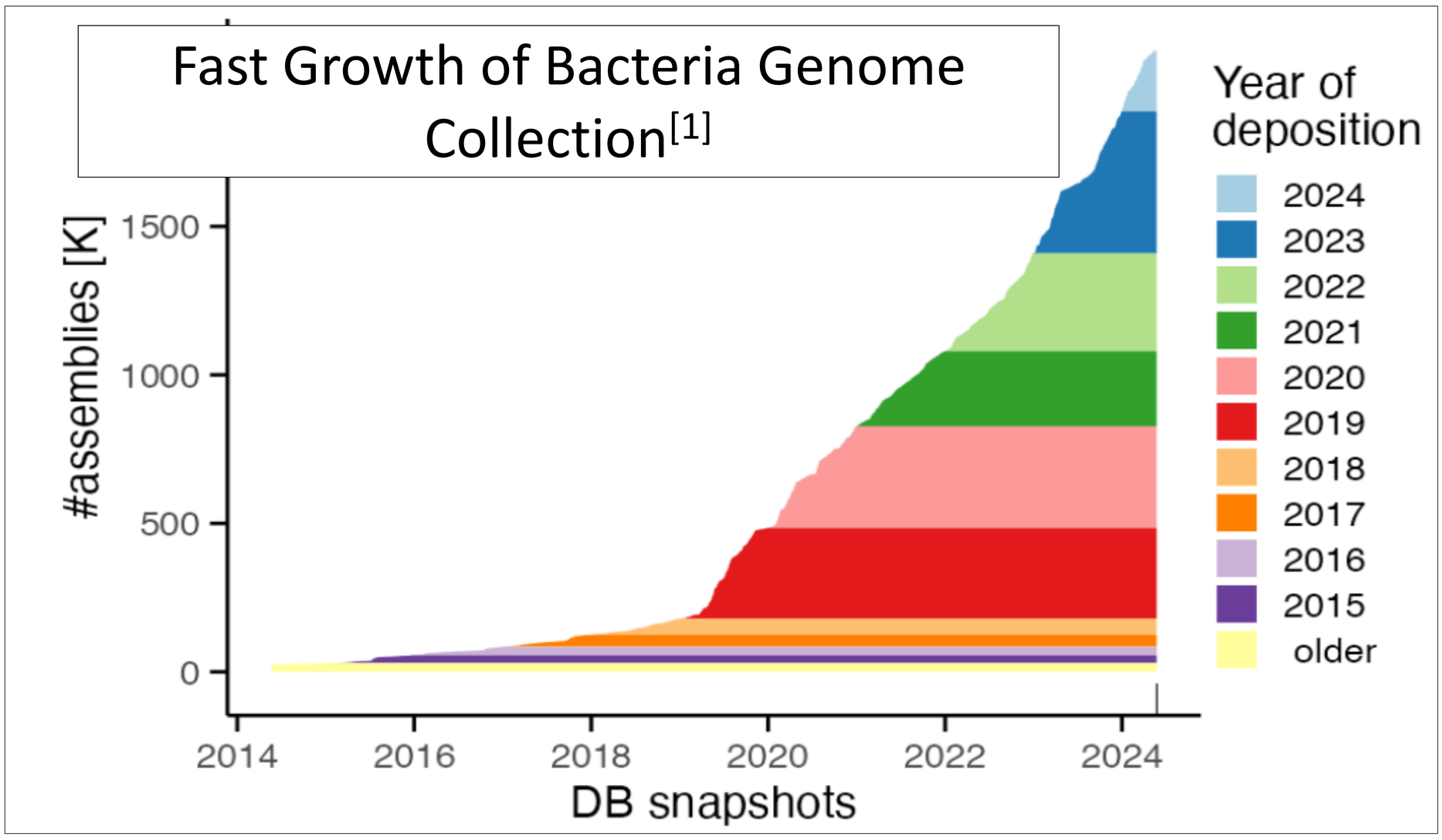
[4] Bonnie et al., DandD: Efficient measurement of sequence growth and similarity. *iScience* 27, 3. 2024

[5] Baker, D.N., Langmead, B. Dashing: fast and accurate genomic distances with HyperLogLog. *Genome Biol* 20, 265. 2019.

[6] Mertens, Stephan, The Easiest Hard Problem: Number Partitioning, in Allon Percus; Gabriel Istrate; Cristopher Moore (eds.), Computational complexity and statistical physics, *Oxford University Press US*, p. 125. 2006

[7] Coffman et al., Bin Packing Approximation Algorithms: Survey and Classification. *Handbook of Combinatorial Optimization* (Vol. 1-5, pp. 455-531). 2012.

Motivation

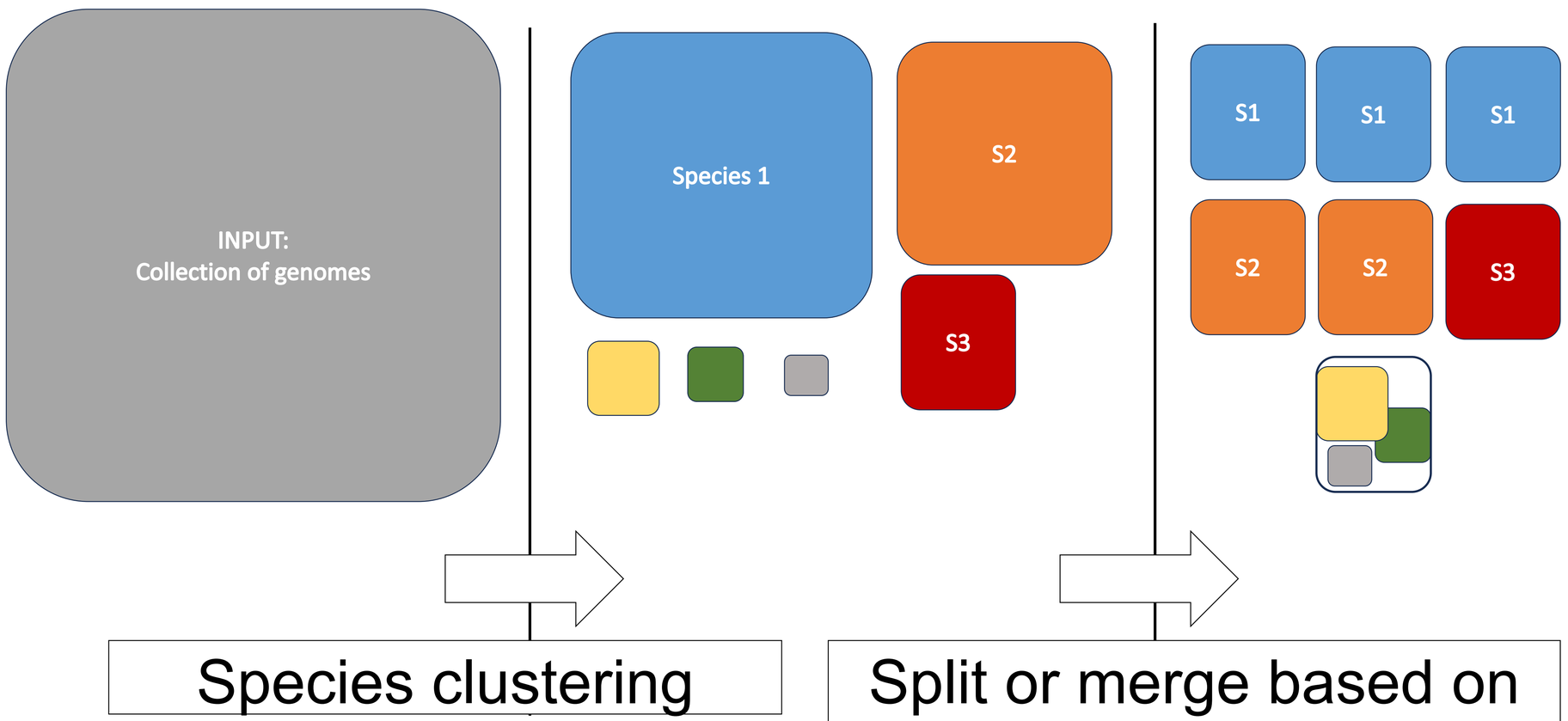


Large Bacterial Genome Collections:	
661k collection[2] (2021)	n = 661,405
AllTheBacteria[3] (2024)	n = 2,440,377
Future	$n > 10^7$

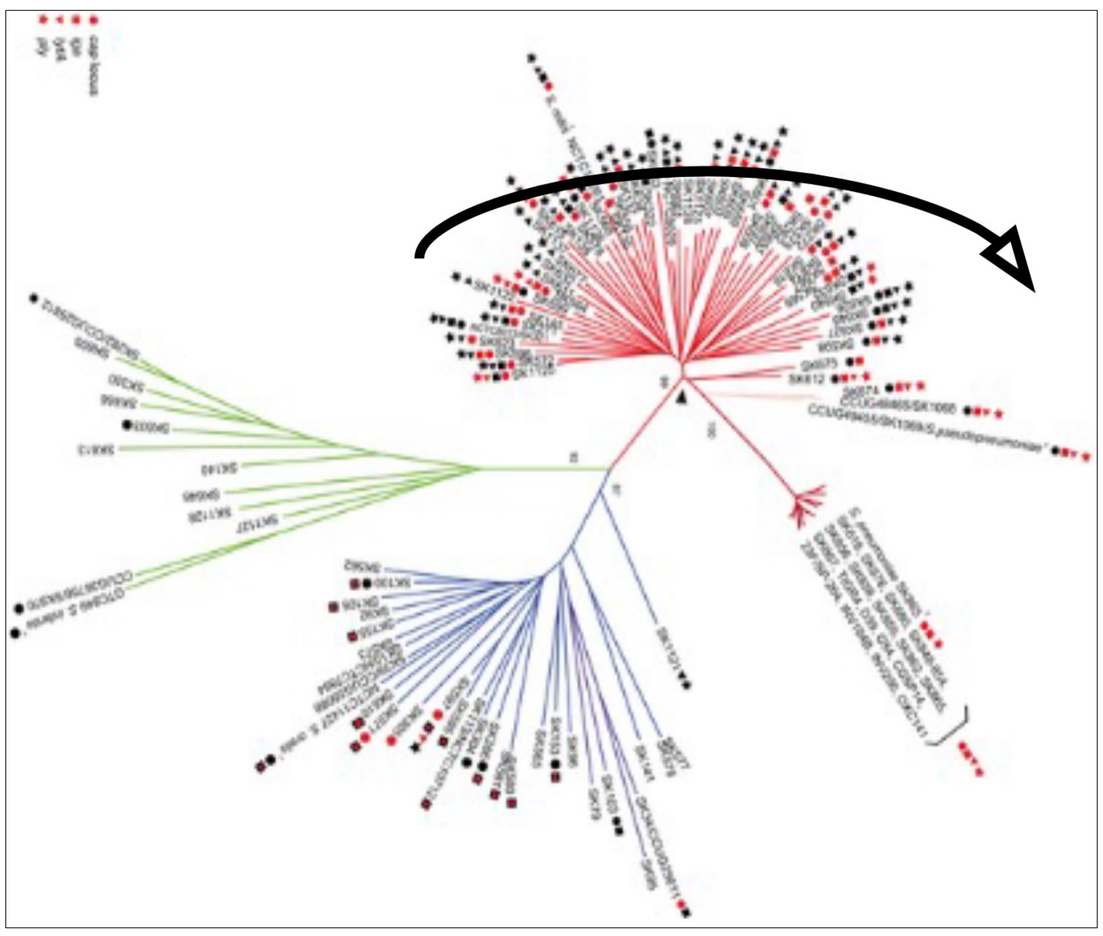
Goal: efficient compression and search within those collections

Recent Innovation: Phylogenetic Compression

STEP 1 : PHYLOGENETIC BATCHING

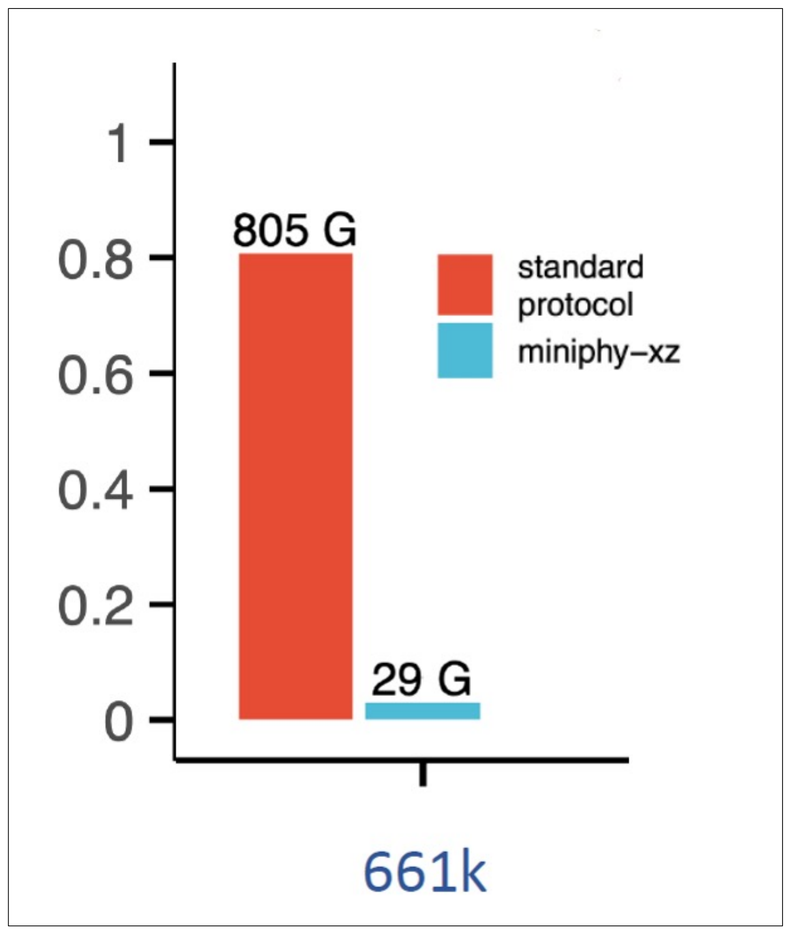


STEP 2: PHYLOGENETIC REORDERING



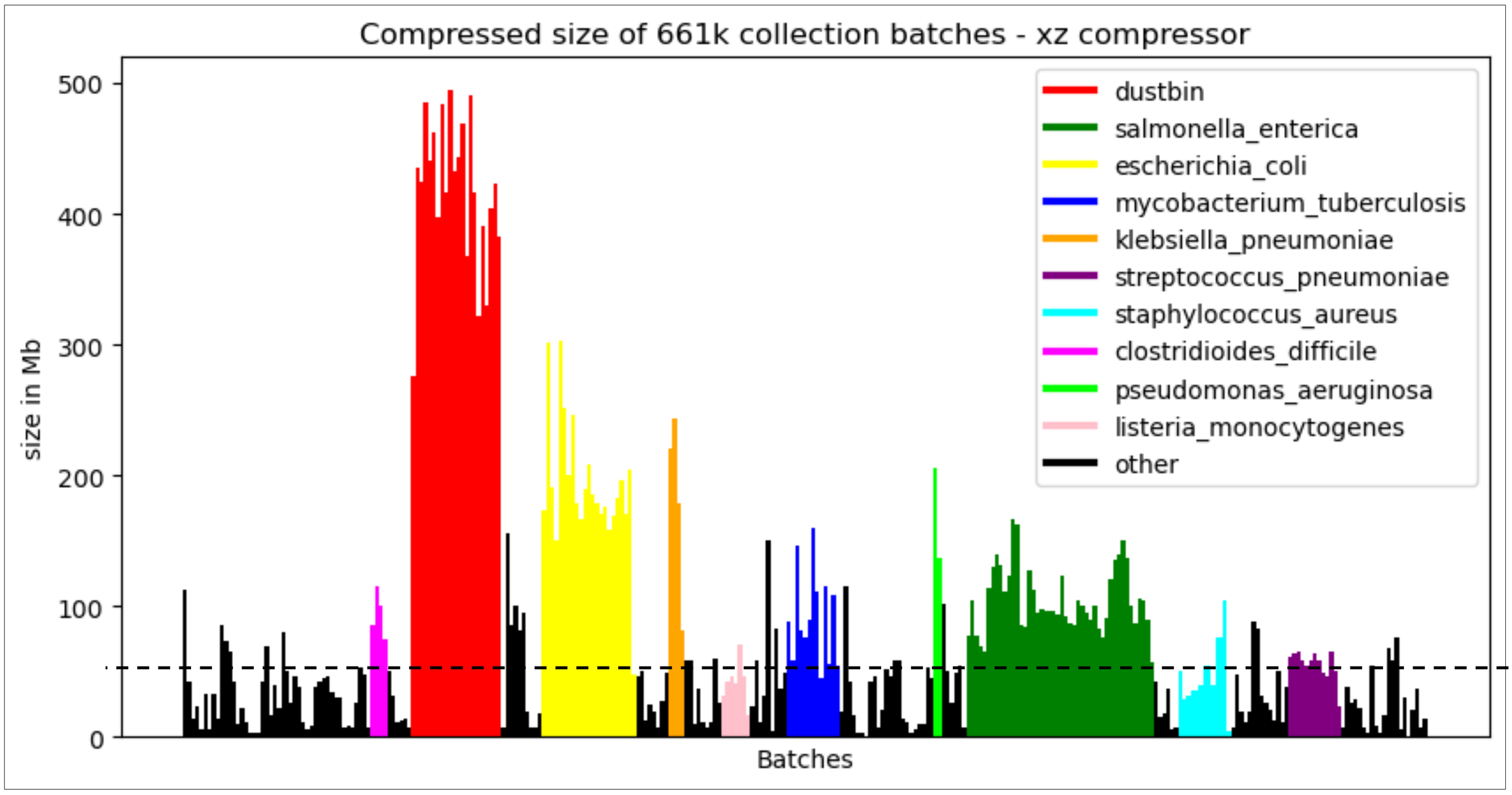
Reordering of genomes in each batch using an estimated evolutionary tree

RESULTING COMPRESSION



Lossless compression of 1-3 orders of magnitude

Current limitation: Batching Results In Non-uniform Compressed Sizes



Unbalanced Workloads

Hinder Parallelization

CONSEQUENCES

Inconsistent Query Times

Memory Overuse

Inefficient Transmission

Ultimate Objective

Objective :

$$min \sum resource(batch)$$

Per-batch Constraints :

Compressed size
Decompressed size
Number of genomes
Search indexes size

Applications:

Portable Devices
(Remote setting, field work, rapid diagnostic)

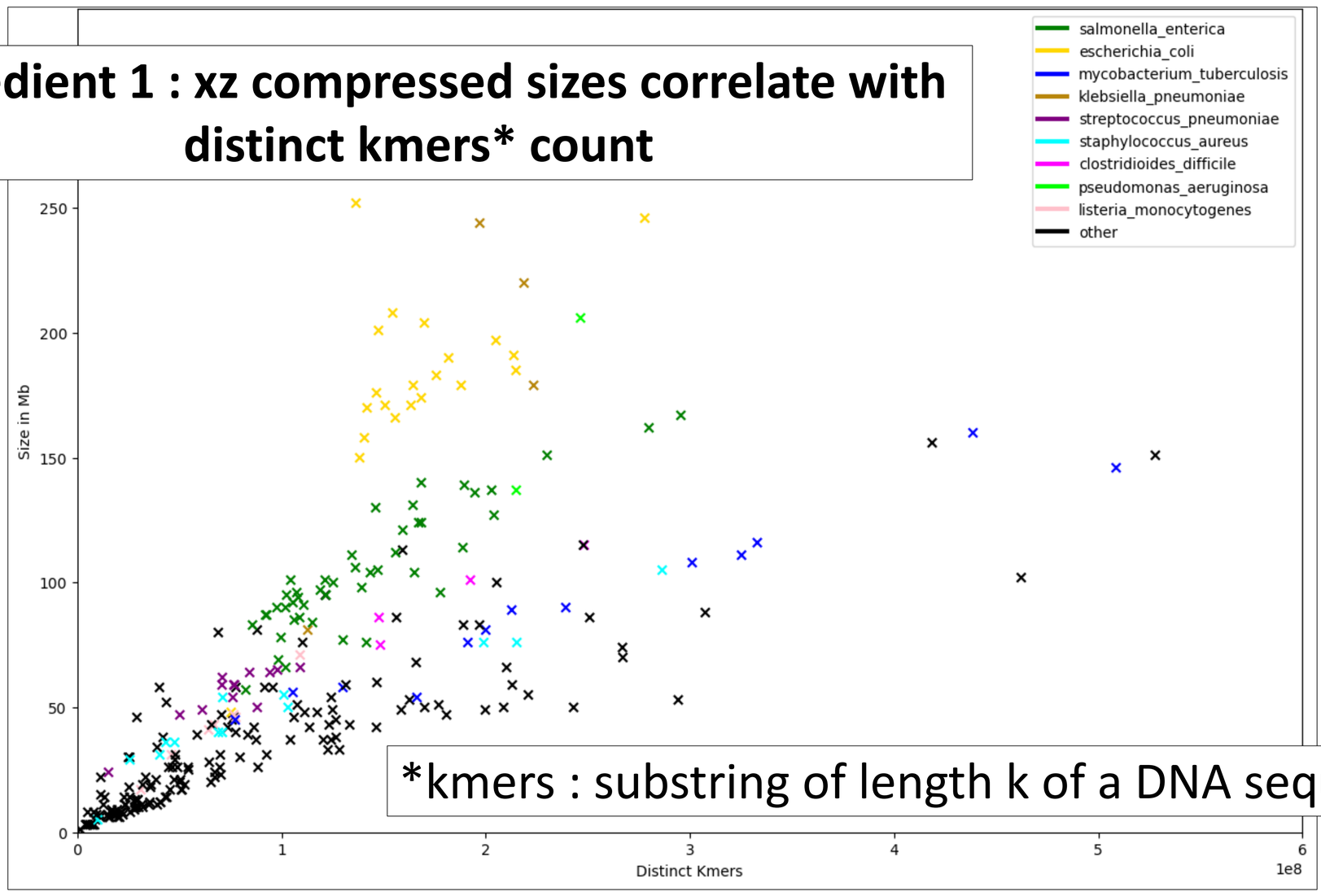
Parallel Platforms
(GPU, Processing-in-Memory)

CURRENT GOAL

Balance post-compression batches for rapid and reliable internet transmission (threshold on post-compression batch size)

Methods

Ingredient 1 : xz compressed sizes correlate with distinct kmers* count



*kmers : substring of length k of a DNA sequence

Ingredient 2: Cardinality estimation using HyperLogLog sketching

Sketches : approximate data structures.

HyperLogLog sketches for cardinality est.: bit patterns,

i.e. *hash(ATGCG)* ➔ 00010100, *hash(CGTAC)* ➔ 00000010.

Fast and efficient UNION operation for sketches.

Preliminary : Given m genomes, put genomes into batches :

STRATEGY 1 : given unlimited batches with capacity C

$$\boldsymbol{Minimize\ nb\ of\ batch\ B}$$

s.t. $distinct\_kmers(bj) < C,\ \text{for (j} = 1, \dots, \text{n)}$

STRATEGY 2 : given a fixed number of batch n

$$\boldsymbol{Minimize\ max\left(distinct\_kmers(b_j)\right), for\ j = 1, \dots, n}$$

Prediction of Genome Batch Post-Compression Size Via Distinct Kmers Estimation

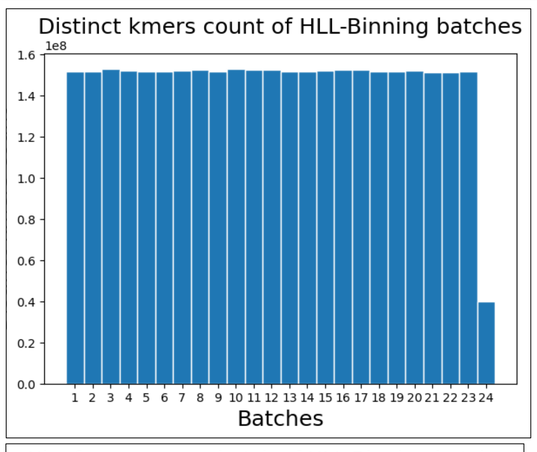
Preliminary results

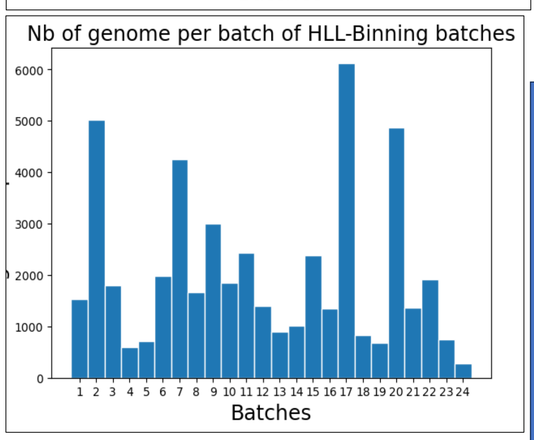
DATA : Genomes of *Mycobacterium tuberculosis* from the 661k Collection[2] , B = 24

STRATEGY 1: HLL-Binning

STRATEGY 2: HLL-Balancing

Batches Obtained From Strat. 1

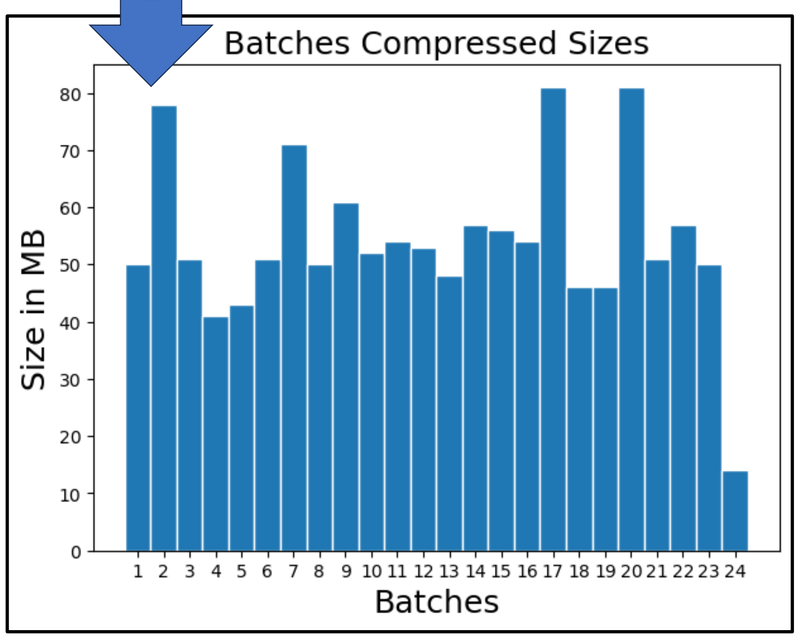




Batch capacity :
C = 152,000,000
(C obtained by linear regression)

Number of genome per batch varies

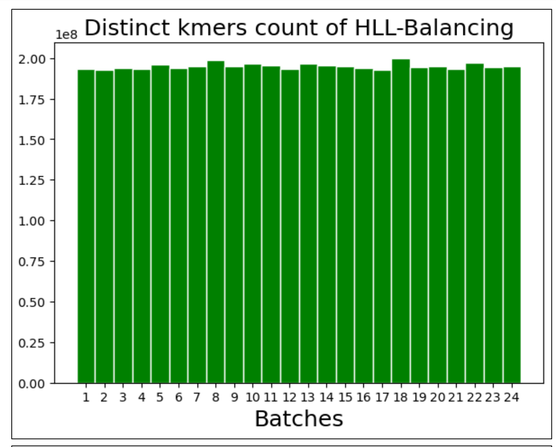
PHYLOGENETIC COMPRESSION

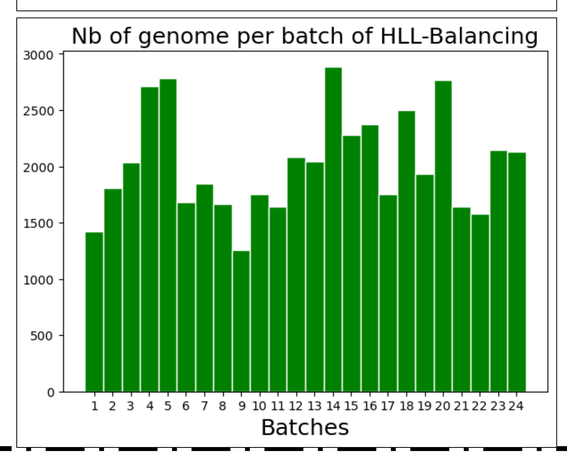


Most of the batches are balanced (between 40-50MB, max size 81MB)

Evaluation strat. 1:
Allowing a capacity on distinct kmers.
The result remains somewhat imbalanced.

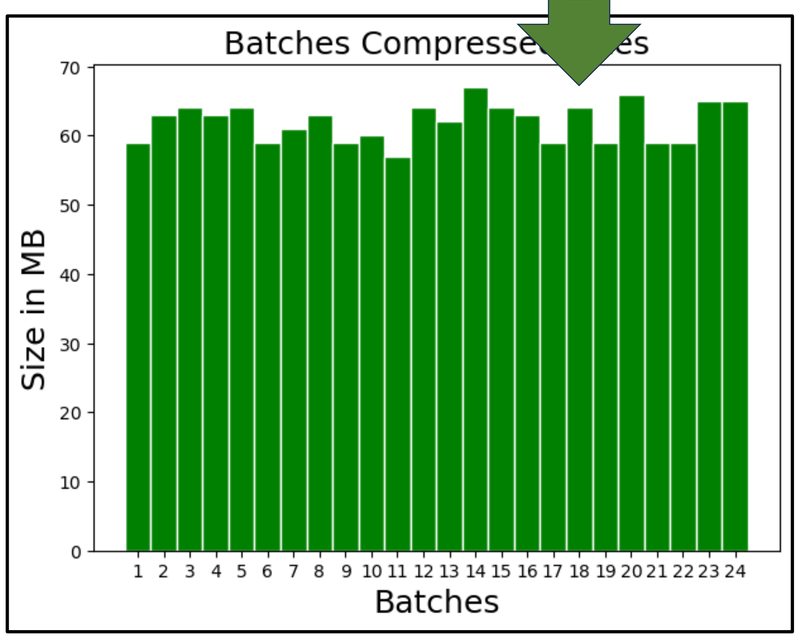
Batches Obtained From Strat. 2





Nb of genomes per batch varies but to a lesser extent compared to Strat. 1

PHYLOGENETIC COMPRESSION



All Batches are well balanced (between 59-67MB, max size 67MB)

Evaluation strat. 2:
Producing more balanced batches.
No control over the maximum distinct k-mer count per batch.

Conclusion and Perspectives

Batching by Predicting Compression Size Using HyperLogLog Distinct K-mer Estimation Improves balancing of the final compressed sizes *Mycobacterium tuberculosis.*

Current Goals:

Extending the results and methods to the whole 661k collection.

Enabling control over the number of genomes in each batch.

Scaling up to AllTheBacteria collection.

Applications in querying data structures such as Bloom filter, on PIM and GPU.

BIBLIOGRAPHY

[1] Břinda et al., Efficient and Robust Search of Microbial Genomes via Phylogenetic Compression. To be appeared in *Nature Methods*. 2025

[2] Blackwell et al., Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences. *PLOS Biology* 19, 11. 2021

[3] Hunt et a.l,. AllTheBacteria - all bacterial genomes assembled, available and searchable. *bioRxiv*. 2024

[4] Bonnie et al., DandD: Efficient measurement of sequence growth and similarity. *iScience* 27, 3. 2024

[5] Baker, D.N., Langmead, B. Dashing: fast and accurate genomic distances with HyperLogLog. *Genome Biol* 20, 265. 2019.

[6] Mertens, Stephan, The Easiest Hard Problem: Number Partitioning, in Allon Percus; Gabriel Istrate; Cristopher Moore (eds.), Computational complexity and statistical physics, *Oxford University Press US*, p. 125. 2006

[7] Coffman et al., Bin Packing Approximation Algorithms: Survey and Classification. *Handbook of Combinatorial Optimization* (Vol. 1-5, pp. 455-531). 2012.