* Need more time on the concept of batches, Rumen should not be confused with this
* “a small subset of species” 🡺 be clear on how many percentages of the database is the small subset. This needs to be measure on 2 cases: pre-compression and post-compression size.
* The concept of capacity needs to be explained better, into more details, we assume compression correlate with the distinct kmers then …
* Strategy 2 not clearly explained, the redistribution of genome needs to be highlighted.
* Strategy 1 would likely to work better
* Take into account high quality vs low quality genomes
  + Do the same experiment for both 🡺 should be know
* When comparing result, always compare the original vs the new one.
* Find a way to compare even though the number of batches are different. Why couldn’t we just change the number of genomes in the original batching strategy.

MAIN POINT:

Improve The Prediction Of The Compression Size.

How: Different tragectories/orders to evaluate comprehensively, for all cases

Orders: Phylogenetic order, accession order, random order.

Write all the questions that I received.

Questions received today:

* How come on the plot some batches have the same number of distincts kmers, but they are not the same in term of compression sizes? (Pierre)
* Is batch the same as bin? (Rumen)
* What is distinct kmers count? (Jacque)
* Why are the new strategy better? Is there an example where the balanceness of batches has a great impact on time performance? (Leo)

Check different statistics for compressors in the Phylogenetic paper

Check AGC by Heng Li