The fast-growing number of available prokaryotic genomes, along with their uneven taxonomic distribution, is a problem when trying to assemble broadly sampled genome sets for phylogenomics and comparative genomics. Indeed, most of the new genomes belong to the same subset of hyper-sampled phyla, such as Proteobacteria and Firmicutes, or even to single species, such as *Escherichia coli* (>3000 genomes as of March 2017), while the continuous flow of newly discovered phyla prompts for regular updates of in-house databases. This situation makes it difficult to maintain sets of representative genomes combining lesser known phyla, for which only few species are available, and sound subsets of highly abundant phyla. An automated method is required but none are publicly available. In this work, the kmer composition of DNA sequences, in conjunction with quality metrics for publicly available assemblies, was used to develop an automated approach for selecting a high-quality subset of representative genomes without redundancy by using our hybrid divide-and-conquer / greedy clustering method.

**Conclusion**

The kmer composition of the genomes can be used to cluster genomes efficiently and has enough signal to avoid taxonomic mixing within the clusters. The method is fast enough to be usable on the evergrowing numbers prokaryotic genomes. The best kmer size so far is size 11, since, with this one only, we can have (1) a very efficient selection in terms of eliminating redundancy at the cost of diversity, (2) a selection maximizing the diversity and avoiding mixing and (3) a selection maximizing the diversity and with a better elimination of redundancy at the cost of a minimal amount of mixing. With only 1 kmer size, we can also reduce the resources needed by having to compute the kmer composition for only 1 size.