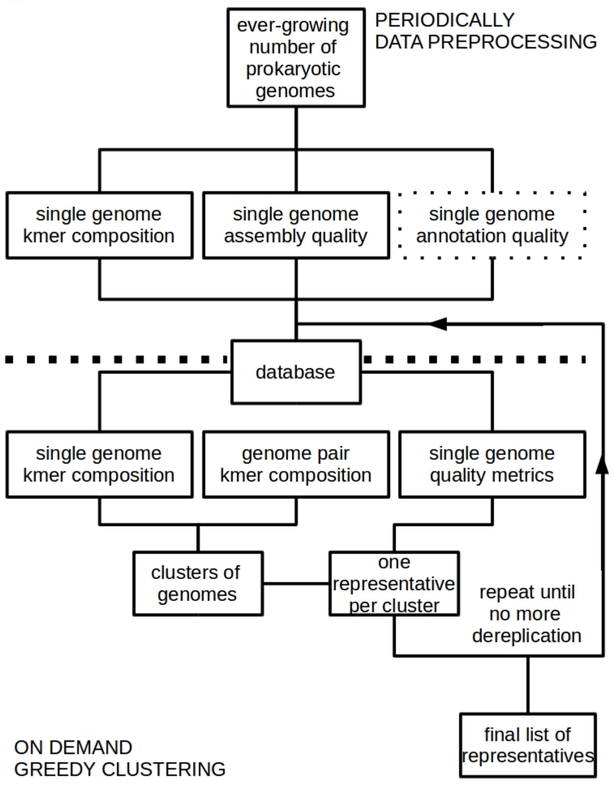
ToRQuEMaDA: Tool for retrieving queried eubacteria, metadata and dereplicating assemblies

Raphaël R. Léonard[1,2*], *Damien Sirjacobs*[2], *Eric Sauvage*[1], *Frédéric Kerff*[1], *Denis Baurain*[2].
[1] InBioS - Centre for Protein Engineering, University of Liège, Belgium
[2] InBioS - PhytoSYSTEMS, University of Liège, Belgium
* rleonard@doct.ulg.ac.be

Introduction

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.

F1

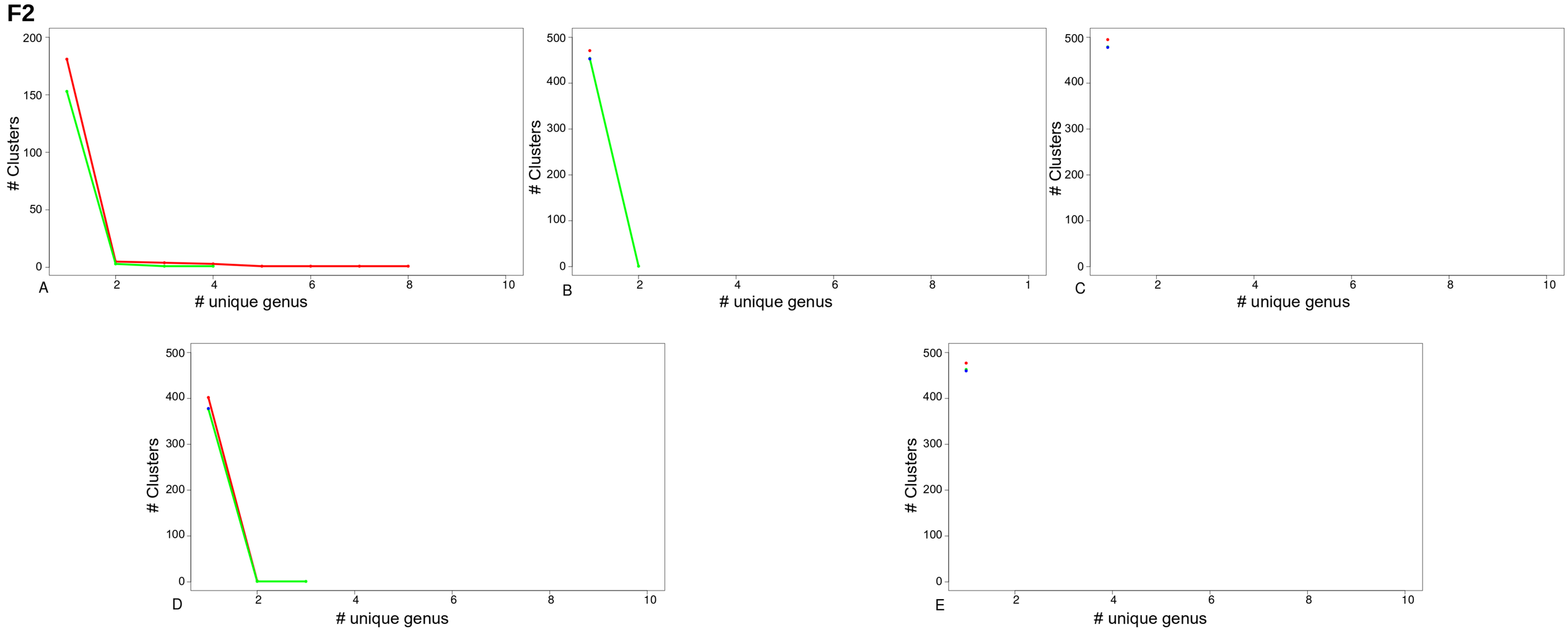


t0.33	base	k11	k13	k15	k17	k19	k21	k11t0.45	k11t0.55
archaea	707	159	454	479	485	487	492	379	461
phyla	9	8	9	9	9	9	9	9	9
class	12	11	12	12	12	12	12	12	12
order	23	20	23	23	23	23	23	23	23
family	41	31	41	41	41	41	41	41	41
genus	120	61	120	120	120	120	120	120	120
reduction	/	+50%	+15%	+12%	+12%	+12%	+12%	+18%	+12%

***T1**: table summarizing the diversity and the redundancy state for 5 kmer sizes at a threshold of 0.33 and also at 2 other thresholds (0.45 and 0.55) for kmer size 11. The threshold is the minimum value for the similarity (Jaccard index) between 2 genomes to be clustered together. The table compares the count of different taxa for a given taxonomic level between the input of ToRQuEMaDA [abbr: TRQEMDA](base) and its output. The reduction row shows the minimal redundancy reduction (of the 3 main phyla) between the different runs of TRQEMDA.*

***F1**: summary of TRQEMDA performance.*

F2



***F2**: the tests were conducted on the Archaea (707 genomes as of March 2017). We tested 6 different kmer sizes for the diversity, the redundancy and the mixing of the clusters (see Methods). Then we tested further 3 of the 6 kmer sizes. The metric used is the taxonomic purity for three different kmer sizes used by TRQEMDA. For each cluster, we retrieve the taxonomy of every member genome and count the number of different taxa observed at the genus level. The plots give the distribution of these counts for the different rounds of TRQEMDA necessary to complete the clustering and for each kmer size. A: kmer size 11, 2 rounds, threshold 0.33 (first red, second green). B: kmer size 13, 3 rounds, threshold 0.33 (third blue). C: kmer size 15, 3 rounds, threshold 0.33. D: kmer size 11, 3 rounds, threshold 0.45. E: kmer size 11, 3 rounds, threshold 0.55.*

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.

Ph.D. Student Day 2017, University of Liege, 11 May 2017

Raphaël R. Leonard is a Ph.D. student funded by the F.R.S.-FNRS.