# Using Dirichlet process to recover genomes from multiple metagenomic datasets

Metagenomics is the study of [genetic](https://en.wikipedia.org/wiki/Genetics) material recovered directly from [environmental](https://en.wikipedia.org/wiki/Natural_environment) samples. Metagenomic samples can contain reads from a huge number of organisms. For example, in a single gram of soil, there can be up to 18000 different types of organisms, each with its own genome. The recovery of genomes from metagenomic datasets is a critical step to defining the functional roles of the underlying uncultivated populations. Here, we deploy a method from the statistics literature called Dirichlet Process; a modeling technique which allows great flexibility in cases where the number of clusters is unknown. While our method uses the same sources of information and models as the majority of algorithms devised for the same problem, it performs better in terms of identifying the number of present genomes and in metrics that asses the quality of all similar clustering procedures resulted from different pipelines.