4. Clustering of RNA-seq reads without reference transcriptome

The goal of *de novo* transcriptomics is to determine the expressed transcripts, their relative abundance levels, and to compare these across conditions and phenotypes when we don't have access to a reference genome. In fact, we don't have a reference genome for most of the species we're interested in studying, and so *de novo* transcriptomics lets us ask the questions we want directly without investing the time / money to first assemble a reference genome. Unfortunately, *de novo* transcriptome assembly is difficult.

Most approaches for *de novo* transcriptome assembly first build a complete de Bruijn graph over the set of reads, which is computationally expensive process. Ultimately, however, related transcripts will be called from related, nearby, and overlapping paths from the same small region of the de Bruijn graph. Thus, instead of building the de Bruijn graph over all sequences reads, we might want to first cluster the reads, and then make a de Bruijn graph on each cluster.

Given the paired end read files we want to find "similar" reads — those which might come from the same "splice graph". There are numerous approaches to this problem which range from ideas based on locality sensitive hashing (i.e. bucket similar reads together) to very recent ideas used to partition reads for metagenomic assembly (a related, but different problem).

Assessment of the quality of the clusterings produced will be done on synthetic data, and in an organism where a reference genome is available. This will allow us to assess the quality of the clusters discovered in a scenario where we know the ground truth (i.e. read that are drawn from the same region of the genome, and thus, belong in the same cluster).