# COMP90016 Workshop 3

Take a closer look at read file reads.fa, which is the same data from the group assignment. You can access it from /home/subjects/comp90016/assignments/group_assignment/.

Q1. If you were to assemble the reads by overlapping them with each other and extending as described in the lecture, what length of the overlapping region would be a good choice? Even if we don't know the length of the reference genome, we can make some arguments in this regard.

Q2. Write a program that counts the number of occurrences of each k-mer for k in {3, 4, 5}. Should the reverse complement of each k-mer be included as well? Discuss in groups in the tutorial.
Given the number of occurrences, which of these k might be used to establish significant overlap? Explain why.

Q3. Write a program that calculates the overlap of length 5 between any two reads. Note that overlaps can happen on both strands. Print all the possible overlaps of all reads. Discuss if this set of reads could be assembled by overlaps based on your evidence. In what ways could the data be improved?