1 Genome Assembly

If you have data from paired-end sequencing, you can will have two separate files from both pairs. You can combine them in an interleaved format.

PacBio has developed a method to reduce their error rate significantly by circularizing the molecule and reading it many times over. They call this circular concensus sequencing.

Before assembling a genome, we need to do some pre processing. First trim reads with low quality calls (towards the end of the reads, it gets worse). Remove very short reads. Correct errors:

- Find all distinct k-mers
- Plot coverage distribution
- Correct low coverage k-mers

To allow for mismatches in an assembly algorithm, you can use MinHash to measure similarity.

JELLYFISH is a good program for generating k-mers.

In real De Bruijn assemblers, k-1-mers are nodes and k-mers are nodes.