Concept 1

MAIN IS NOT TRUE

1.1 Pacbio reads

Daligner finds alignments between long, noisy reads. Pacific Biosciences has commercially launched its first sequencer in 2011. It is able to output reads with an average of 1000 bases, which is significantly longer than NGS reads [**?**]. In 2014, a new polymerase-chemistry combination was released, called P6-C4. This version can output average read lengths of 10000-15000 bases, and its longest reads can exceed 40000 bases [**?**]. While the drawback is that these reads have an error rate of 12-15%, this can be compensated by the distribution of these errors [**?**]. First, the set of reads is a nearly Poisson sampling of the sampled genome. This implies that there exists a coverage c for every target coverage k, such that every region of the genome is covered k times [**?**]. Secondly, the work of Churchill and Waterman [**?**] implies that the accuracy of the consensus sequence of k sequences is $O(\epsilon^k)$, which goes to 0 as k increases. This means that if the reads are long enough to handle repetitive regions, in principle a near perfect de novo assembly of the genome is possible, given enough coverage.

Important points for de novo DNA sequencing are: what level of coverage is needed for high quality assembly? And how to build an assembler that is able to deal with high error rates and long reads? Most previous assemblers work with NGS reads, which are much shorter and have much lower error rates. Some algorithms used in these assemblers, such as de-Bruijn graphs [**?**] would fail

However, since Daligner was build, new methods of using de Bruijn graphs with long reads have been developed.