HIFI reads

Until lately researchers had to rely on low error short reads and error prone long read sequencing. Both technologies had certain advantages and disadvantages. The short reads had a high accuracy but which are usually only about 150bp long. This makes it hard to resolve certain regions, which do not contain much variation, like repeats, leaving many short read assemblies fragmented (e.g. according to de Koning et al. (2011) up to approx. 2/3 of the human genome might is repetitive). Here, long read sequencing technology has its advantages, with reads usually being 10 – 100kb long. However, the high error rate of up to 15~% represents a problem. Usually a genome is assembled using long reads and is then polished using the low error rate of short reads. In simplified terms, the short reads are simply mapped against the error prone assembly and sites at which the assembly does not match the short read data are corrected.

The latest development in sequencing technology are HIFI- or Circular Consensus (CCS)-reads. This technology tries to combine an error rate similar to the one of short reads with the read length of long reads, making it easier to assemble and analyze genomic data (Fig. 1).

Figure 1: HIFI-reads have a similar read length as usual Pacbio reads, but have a much higher accuracy.

The idea behind HIFI-reads is fairly simple. An adapter is added to the DNA-molecules, which circularizes them. This way it is possible for the polymerase to loop around the DNA-molecule several times, each time generating an error prone, "normal" long read, so called sub-reads. Because the adapters are also sequenced one can easily tell where a new subread starts and where it ends. This way the subreads of the two DNA-strands can be compared and a consensus sequence for each read can be built. This way random sequencing errors cancel each other out, resulting in high accuracy long reads (Fig. 2). However, this technology is still quite expensive, since a lot of sequencing data is required. For example the approx. 9Gb of HIFI data used in the course were generated from approx. 560Gb of raw data.

Figure 2: The high accuracy of HIFI-reads is reached by first circularizing the DNA using adapters and then sequencing the same DNA-molecule several times. Using the adapters one can easily tell where one so called subread starts and where it ends in the large polymerase read. These subreads can be compared to each other and used to build a consensus sequence, in which random sequencing errors cancel each other out. This Circular Concensus Sequence (CCS) has a sequencing error in the range of what is seen for the high quality Illumina reads (<0.1%).

From:
https://applbio.biologie.uni-frankfurt.de/teaching/wiki/ - **Teaching**

Permanent link:
https://applbio.biologie.uni-frankfurt.de/teaching/wiki/doku.php?id=general:bioseqanalysis:hifireads

Last update: **2021/10/21 11:12**