DNA sequencing

Methodologically, the sequencing of DNA is straightforward. Most technologies are based on an enzymatic sequencing by synthesis. Here, modified DNA polymerases from thermophilic bacteria synthesize DNA in vitro using the genomic DNA of the target organism as template. The methods then differ, in part substantially, in the way of how they detect the incorporated nucleotide, and by doing so, determine the sequence of the DNA. A recent poster summarizing the 'evolution of sequencing technology ' provides a concise entry into the field. Common to all contemporary sequencing methods is, however, that the length of genomic DNA that can be consecutively sequenced, often referred to as the read length, is tiny to small compared to the genome size. Typical read lengths range from 75 base pairs (bp) up to a few thousand bp, depending on the technology. Figure 1 provides an overview of the existing technologies together with their average read lengths and output per run.

Figure 1: Overview of the existing sequencing technologies. The figure reveals that sequencing technologies either optimize for throughput (Illumina HiSeq) or for read length (e.g. Oxford Nanopore).

Over the years, the 3rd generation sequencing technologies have improved dramatically both with respect to read length and data quality. See for example the publication Opportunities and challenges in long-read sequencing data analysis by Amarasinghe et al. (2020). A major improvement were the development of the PacBio HiFi protocol that decreased the sequencing error of PacBio reads from around 15% down to less than 1%. This resembles almost the sequencing error of Illumina reads, however with the advantage of read lengths up to 20,000 - 30,000 bp.

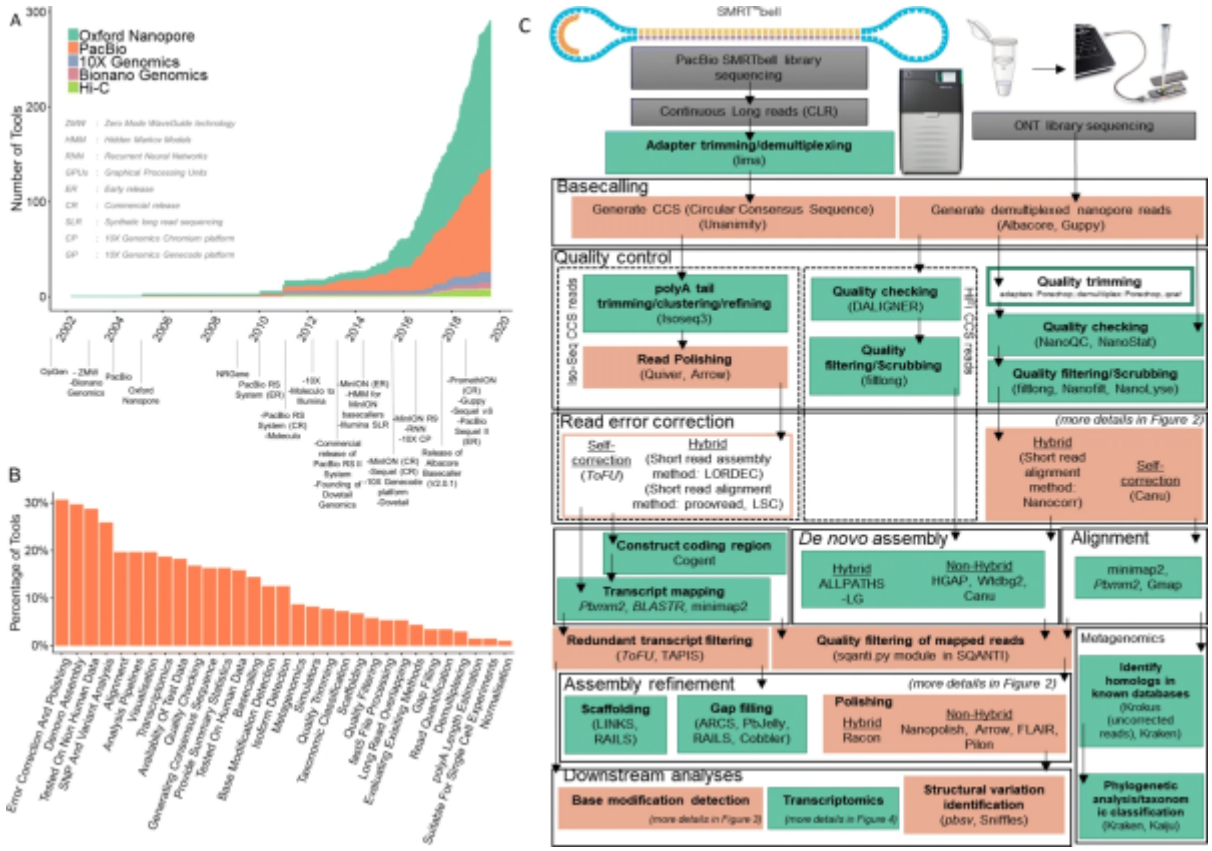

Figure 2: Overview of long-read analysis tools and pipelines. a Release of tools identified from various sources and milestones of long-read sequencing. b Functional categories. c Typical long-read analysis pipelines for SMRT and nanopore data. Six main stages are identified through the presented workflow (i.e. basecalling, quality control, read error correction, assembly/alignment, assembly refinement, and downstream analyses). The green-coloured boxes represent processes common to both short-read and long-read analyses. The orange-coloured boxes represent the processes unique to long-read analyses.

Unfilled boxes represent optional steps. Commonly used tools for each step in long-read analysis are within brackets. Italics signify tools developed by either PacBio or ONT companies, and non-italics signify tools developed by external parties. Arrows represent the direction of the workflow. Figure from Amarasinghe et al. 2020

- Read more about PacBio HiFi reads
- Read more about Shotgun sequencing