Related Work

The process of de-novo assembly has to do with assembling a species genome for the first time. The resulting genome can then be used to facilitate genome assembly of other individuals of the same species, in a process called the reference-guided assembly. De-novo assembly refers to assembling the genome without having a draft genome; this is a complex and difficult problem to solve. This leads to a less accurate assembled genome, we can consider only exact-match overlaps instead of all overlaps between reads. Scalability is another expectation from an assembly pipeline. Genome assembly requires a massive amount of processing, and a pipeline should be able to utilize all available hardware resources such as processors and memory with the highest efficiency. The program should also be able to deal with limitations. For instance, using a large amount of memory is common in de-novo assembly programs. A scalable program should be able to deal with the limited amount of available memory, with minimal impact on execution time. In addition to the above expectations, the choice of an assembly pipeline should be based on the following 'data-specific parameters.

- In a recent review

1- Sohn et al. compared several short-read, long-read and hybrid assembly pipelines. However, they did not perform their own analysis; rather, they used the results provided by the author of each method for the evaluation. Thus, their comparison metrics were limited to N50, which is provided by all authors.
2- In Bradnam et al. evaluated multiple assemblies submitted by several participating teams. Their main focus was on short-read assembly. Each participating team best optimized the pipeline for the given dataset. However, the error pattern varied in accordance with the sequencing technology and sample preparation (i.e., contamination). Participating teams were able to evaluate their assembly using closely related genomes prior to submission.
3- The work in is another review paper in which long-read sequenced data (Pac Bio) are not considered. Neither of these surveys includes the most recent long-read assemblers.

In this review, we have discussed the challenges in de novo whole-genome assembly for short reads and how these challenges can be overcome. Notably, the methods using short reads perform poorly for repetitive structures or GC-biased regions. Assemblers using long reads are highly helpful in resolving the problems, and they generate more contiguous results. However, the sequencing and computational costs are currently substantially higher for long reads. At a minimum, the sequencing cost problem can be mitigated by the hybrid approaches of short NGS reads and long reads, although there is an argument that the hybrid approach may produce numerous misassemblies .

The scaffold size should be close to the chromosome scale with a low number of gaps to ultimately generate the de novo assembly. Recently, chromosome-scale scaffolding has been achieved by using

advanced physical mapping methods, such as optimal mapping and chromatin-interaction mapping. Optimal mapping takes advantage of genome-wide restriction site maps from a single DNA, generating ordered, high-resolution restriction maps. In contrast, genome-wide chromatin interaction mapping provides ordered information about the long-range interactions within a chromosome, including centromeres, and was originally designed to detect the regulation of gene expression by the three-dimensional interactions of chromosomes. Recently, this method has been applied to ultra-long-range scaffolding of de novo assemblies, building highly qualitative chromosome-scale scaffolds. Given the successful scaffolding methods, the advent of new technologies for generating genome-wide, ultra-long-range physical maps should significantly improve the quality of the de novo assemblies in the near future.

Future sequencing technologies may also offer improvements. Currently, the original PacBio long read sequencing is expensive. However, with the advent of Oxford Nanopore and PacBio sequel platforms inexpensive long read sequencing technologies are within reach and may lead to long-read-only assembly being more frequently used. It has been also expected that quantum sequencing technologies may further reduce the cost problem by increasing the throughput and read length . According to Di Ventra and Taniguchi, the throughput of quantum sequencing.Although this throughput is only an expectation, it appears promising that the throughput problem of long SMS reads can be solved. Once the high-throughput sequencing of long reads becomes a reality, the current long read assemblers would be not suitable, owing to overflowing memory, and thus a high-priority challenge in de novo assembly will be the development of new assembly algorithms with efficient memory and computational costs.

The Oxford Nanopore Company provides an online-based platform for real-time data analysis. The real-time assemblies for large genomes are not currently available, although assemblies for the small bacterial and eukaryotic genomes are available. For the real-time assembly of large genomes, the error correction and assembly algorithms for erroneous long reads must be much more efficient than ever before. In particular, the Oxford Nanopore reads often include indels, and the indel-mediated mismatches may be abundantly detected during read overlapping and error correction, thus potentially resulting in error-prone assembly. Methods for addressing these problems must be developed in the future.

In addition, as interest in precision medicine and personalized genomics increases, scientists and nonscientists are expecting to analyze their genomes to understand their current and future health conditions on a personal computer or smartphone. Genome analyzers that include assembly may need to be implemented as a light application that requires a small amount of memory and computing to make precision, personalized medicine a reality.

Together, the data suggest that given the advanced technologies and clinical interests in personal genomes, more memory- and computing-efficient technologies to generate de novo assembly of personal genomes and metagenomics will be beneficial for clinical uses in the future.