Will long read sequencing technologies replace short read sequencing technologies in the next ten years?

Nishat Anjum Bristy

March 6, 2020

Accurate and comprehensive identification of genetic variation is extremely important for determining the genetic basis of diseases and expression of characteristics. Before the advent of long read sequencing, high-throughput short read sequencing has been very successful in assessing the genetic variations due to single nucleotide polymorphisms (SNPs). But a large amount of genetic mutations such as insertion, deletion, duplication, translocation, repetition were undetected due to the length of reads being short. De novo assembly of high coverage third-generation long reads has enabled us to achieve highly contiguous, continuous and complete genome sequence. So, at the very first glance, it might seem like long read sequencing technologies will prevail over short read sequencing but whether to use long read or short read sequencing totally depends on the researcher, biologist or scientist's choice and requirement. In my opinion, sequencing is far from a "Whether-Or" idea. So, whether long read sequencing will replace the short read sequencing within the next 10 years is also a relative question. In medical genomics and disease research, where genomes tend to have a lot of mutations, structural variations, long read sequencing will rapidly take over short read sequencing. But still there will be many cases which will need affirmations from short read sequencing.

Sequencing is one of the fundamental component in a large number of fields like molecular biology, evolutionary biology, metagenomics, virology, medicine, forensics and so on. Sequencing can be used for individual genes, larger genetic region, whole chromosomes or even the entire genome of any organelle. In broader sense, three types of sequencing have been in practice - sanger sequencing, high throughput short read sequencing and recently developed third generation long read sequencing. Sanger sequencing uses chain termination method and short read sequencing uses the shotgun sequencing method which aims at sequencing an entire chromosome, long DNA or RNA fragments with more than 1000 bases. Shotgun sequencing is one of the pioneering technologies which enabled the whole genome sequencing at the first place. High throughput short read sequencing method is also known as next generation sequencing (NGS). NGS is different from sanger sequencing in that it provides massively parallel analysis at a much reduced cost. The cost of sanger sequencing is $\$500$ per 1000 bases whereas NGS has reduced the cost by its massively parallel nature to $\$0.5$ per 1000 bases. So, it is evident that short read-sequencing technologies are cost effective, reliable and there is a large number of tools that support NGS. But third generation sequencing technologies emerged in the last five years are also coming to the brink of maturity with approaches to error rate minimization and cost optimization.

Although the cost of NGS is very economical in comparison to other technologies, this is not the only case to be considered for the application of sequencing. Human genome is over 3 billion base pairs long. Short read sequencers such as Illumina's NovaSeq, MiSeq, HiSeq, BGI's MGISEQ, BGISEQ, or Thermo Fisher's Ion Torrent sequences [1] can sequence upto 600 bases per read. Among all the long read sequencing technologies, two most common are Pacific Bioscience's (PacBio) single-molecule real-time (SMRT) and Oxford Nanopore Technologies' (ONT) nanopore sequencing. 10X and Dovetail technologies are still under development. SMRT can sequence from 250 bp to 50 kbp and and nanopore sequencing can sequence from 500 bp to 2.5Mbp [2]. It is evident that extremely high parallel processing power is needed for short read sequencing. Whereas, long read sequencing gives benefit in this case through its longer reads. Moreover, in short read sequencing, reads are assembled into contigs and then contigs are joined into scaffolds. Scaffolds contain gaps composed of repeats which are masked while scaffolding. Without longer reads, repeats that extend beyond the paired-end short reads will not be assembled. Long read sequencing improves de novo assembly by constructing longer contigs, and thus is able to detect transcript isoforms, sequence regions with extreme GC content bias and structural variations such as repetitive regions, insertion, deletion or transposition.

However, this is to be stated that the rapid advancement of the sequencing field occurred due to the emergence of NGS technologies. In the past 10 years, the wide implementation of NGS has fundamentally changed the field of medical genetics, plant biology, disease detection, drug preparation and so on. But even after using very sophisticated bioinformatics algorithms for assembling, short read sequencing technologies are limited in identifying several types

of structural variations. A recent de novo assembly of a *Drosophila melanogaster* reference genome has revealed that short read sequencing methods miss hundreds of structural variants which affect the characteristics of various important phenotypes [3]. This fact is alarming because the limitations of short reads can contribute to incorrect diagnostic of genetic disorders in patients. With short read sequencing, data analysis is highly dependent on reference genome and variant phasing information is often lost. Dependence on reference genomes can lead to wrong judgement about genetic diseases because many structural variations and mutations occur in the case of such genetic diseases [4]. However, long read sequencing not only has the potential to produce very high quality genome sequences but also has the ability to identify various critical genetic diseases and mutations. Oxford nanopore sequencers with MinION, GridION, PromethION are the most impressive among the long read sequencing technologies. A single strand of DNA molecule is passed through a nanopore which is inserted into a membrane, with an attached enzyme. Changes in the electric signal is measured and the base is identified. Structural variations which are greater than 50 bps long causes significant genetic variation. SMRT sequencing of a haploid human genome has shown that almost 89% structural variations were missed by the 1000 genome project(launched in 2008 with the motivation of establishing the most detailed human genome) [5]. Moreover, scientists are now able to discover novel disease genes harboring SVs that are sequenced with long read sequencing technologies. Many diseases are caused due to tandem repeat expansions. These disease causing expansions are longer than current NGS short reads, causing incorrect disease detection. SMRT and nanopore long read sequencing has been able to detect these repeat expansions correctly. This suggests that most of the genetic disorders which were difficult to identify require long reads for correct identification of the variations of the population gene from the reference gene. Thus, SMRT and Nanopore long read sequencing offers promising alternatives to high-throughput short read sequencing.

However, the main hindrance in using long read sequencing is its high error rate because human genome is almost 99.9% same for all the individuals. Only 0.01% varies from person to person. So, the high error rate of long read sequencing is a problematic issue that needs proper attention. In comparison to sanger and next generation sequencing, sanger and NGS(Illumina sequencing) has raw error rate of 0.3% and 0.8% while SMRT has an error rate of 12.9% and nanopore sequencing has an error rate of 34% [6]. Some recent studies and tools have already lowered the error rate of nanopore sequencing to 13.72% for raw reads [7]. Most of the error correction tools for long read sequencing such as RAW, HALC(t), HALC(s), LoRDEC, NaS(μ) has been successful in lowering the error upto 0.33% using hybrid methods constituting both short and long reads. Non-hybrid self-correcting tools are also being successful in the recent years. A study showed that de-novo assembly of a microbial genome with self-correcting SMRT sequencing alone gave much better results than short read sequencing [8]. Till now, hybrid tools are the most successful ones because these tools leverage the benefits of both short and long read sequencing.

Despite the high error rate in sequence estimation, the usage of long read sequencing is increasing everyday due to its power of identifying novel type of genetic mutations. Diseases like muscular dystrophy due to tandem repeats, cancer, autism, HIV have higher probability of getting detected and studied because of the advancement in sequencing technologies. Cancer cells possess mutations in their genome which result in the abnormal growth or initiation of cancer by altering the function of driver genes and tumor suppressor genes. With the advancement in sequencing technologies, these mutations are identified and various drugs are designed to target the mutated region. So far short read sequencing has been able to identify point mutations such as single nucleotide variants (SNV) and short indels. However, complex genetic changes like structural variations (SVs) or mutations in repetitive regions are not identified by short reads. Long read sequencing is able to identify complicated structural variations with combinations of local duplication, inversion or deletions in tumor suppressor genes. Moreover, long reads are able to fully cover carcinogenic transcript sequences (also known as fusion transcripts). Structures of these transcripts can be fully determined by long read sequencing of complementary DNA (cDNA) [9].

Long read sequencing also has the potential to detect diseases due to tandem repeats. There are more than 30 diseases which are caused due to tandem repeats. Two most common of them are myotonic dystrophy and facioscapulohumeral muscular dystrophy (FSHD). Sequencing of these diseases is very challenging using traditional sequencing approaches because of their subtelomeric tandem repeats and high GC content. Historically, most of these diseases were identified by linkage analysis in families with multiple affected members. Mitsuhashi et al shows from the Human Genetic Mutation Database (HGMD) that the tandem repeat expansion length tends to be long in untranslated regions and that nanopore or SMRT long read lengths can cover known expansions. This observation indicates that undiscovered repeat diseases within this range of repeat expansion will be found by long read sequencers [10]. Alternative splicing of HIV-1 sequenced by long read sequencers has opened a new window for HIV-1 virus study and drug development. Moreover, a large study on autism showed that many regions with SVs are still undiscovered and short read sequencing technologies are not good enough for detecting these SVs. All these examples support the fact that the field of genetic disease detection is in its booming phase due to long read sequencing technologies.

From the above discussion, it can be clearly seen that long read sequencing technologies have already replaced NGS for complex genetic diseases. It provides crucial information about the genome including structural variations, base modifications, haplotype phasing, transcript isoforms etc. In recent years, the drive for sequencing technologies is towards long read sequencing. However, long read sequencing technologies have several barriers to overcome. It is not always possible to obtain high quality samples of DNA and RNA from medical samples. For this reason, the accuracy of long read sequencing is 90% nowadays which makes it difficult to determine the point mutations. NGS can help in these cases of point mutations and error correction. But due to the importance in identifying the unknown regions of the genome, one of the main issues of bioinformatics nowadays is the development of tools for long read sequencing but it is a matter of sorrow that none are still resilient enough to sustain with the high error rates of raw long reads. Since its introduction in 2008-2009, long read sequencing has already taken the studies of complicated genetic disorders, cancer research, epigenetics to a new era. This field is evolving rapidly and new technologies for error correction and high coverage confirmation are under development.

Hence, I believe that the preparatory stage for long read sequencing technologies has already passed and the next 10 years are going to be the take-off period of this quickly emerging field. In the next 10 years, long read sequencing technologies will have stronger research and commercial grip than ever before but NGS will still be needed for error correction and detection of SNPs. So, this is my observation that long read sequencing technologies are definitely going to take over eventually but in next 10 years, the long and short read sequencing technologies will be used in tandem.

References

[1] Sara Goodwin, John D McPherson, and W Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333, 2016.

[2] Shanika L Amarasinghe, Shian Su, Xueyi Dong, Luke Zappia, Matthew E Ritchie, and Quentin Gouil. Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, 21(1):1–16, 2020.

[3] Edwin A Solares, Mahul Chakraborty, Danny E Miller, Shannon Kalsow, Kate Hall, Anoja G Perera, JJ Emerson, and R Scott Hawley. Rapid low-cost assembly of the drosophila melanogaster reference genome using low-coverage, long-read sequencing. *G3: Genes, Genomes, Genetics*, 8(10):3143–3154, 2018.

[4] Tuomo Mantere, Simone Kersten, and Alexander Hoischen. Long-read sequencing emerging in medical genetics. *Frontiers in Genetics*, 10:426, 2019.

[5] John Huddleston, Mark JP Chaisson, Karyn Meltz Steinberg, Wes Warren, Kendra Hoekzema, David Gordon, Tina A Graves-Lindsay, Katherine M Munson, Zev N Kronenberg, Laura Vives, et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome research*, 27(5):677–685, 2017.

[6] Jerzy K Kulski. Next-generation sequencing—an overview of the history, tools, and "omic" applications. *Next Generation Sequencing–Advances, Applications and Challenges*, pages 3–60, 2016.

[7] Leandro Lima, Camille Marchet, Ségolène Caboche, Corinne Da Silva, Benjamin Istace, Jean-Marc Aury, Hélène Touzet, and Rayan Chikhi. Comparative assessment of long-read error correction software applied to nanopore rna-sequencing data. 2019.

[8] Chen-Shan Chin, David H Alexander, Patrick Marks, Aaron A Klammer, James Drake, Cheryl Heiner, Alicia Clum, Alex Copeland, John Huddleston, Evan E Eichler, et al. Nonhybrid, finished microbial genome assemblies from long-read smrt sequencing data. *Nature methods*, 10(6):563, 2013.

[9] Yoshitaka Sakamoto, Sarun Sereewattanawoot, and Ayako Suzuki. A new era of long-read sequencing for cancer genomics. *Journal of human genetics*, pages 1–8, 2019.

[10] Satomi Mitsuhashi and Naomichi Matsumoto. Long-read sequencing for rare human genetic diseases. *Journal of Human Genetics*, pages 1–9, 2019.

Word count : 1940