Possibility of using long read sequencing for analysing the mRNA transcriptome of cells

Wenfa Ng

Citizen scientist, Singapore, Email: ngwenfa771@hotmail.com

Abstract

Differential gene expression under different environmental and nutritional conditions is one approach where perturbation studies could gain insights into the various stress responses endowed in a particular species for responding to myriad environmental stressors. Such differences in gene expression could manifest at two levels: (i) mRNA transcript, or (ii) protein level. With the advent of high throughput next generation sequencing, it has become easier to probe gene expression at the mRNA level compared to the protein level. Hence, RNA-seq has emerged as the dominant technology for probing differential gene expression. Using short-read sequencing of fragments of mRNA, RNA-seq typically generates a large dataset comprising millions of short reads, which after being mapped to a reference genome, could provide a quantitative assessment of relative expression levels of different genes under different conditions. However, one problem with current RNA-seq methodologies lies in the fragmentation of long mRNA transcripts that comprise various gene expression control elements and the coding gene into short fragments. Fragmentation of mRNA transcripts reduce the biological information that could be harnessed from the dataset. For example, reducing the length of the accessible mRNA transcript to segments of 50 nucleotides would not provide additional information on the 5' untranslated region such as the ribosome binding site, even though 50 nucleotides provide sufficient biological information for mapping the sequenced reads to a coding sequence in the genome. More importantly, fragmenting a long mRNA transcript into smaller fragments may artificially inflate the read counts during mapping of sequenced reads to the reference genome. Specifically, a single mRNA transcript could yield multiple short fragments that could be mapped to the same coding sequence on the reference genome; thereby, inflating the read counts. To ameliorate this important problem, long read sequencing such as those available from Pacific Biosciences and Nanopore sequencing could be applied to sequence the mRNA library of a cell. No fragmentation of the mRNA transcripts would be induced, which would help solve the read count inflation problem described above. Hence, application of long read sequencing technologies to access the transcriptome of a cell would provide manifold advantages such as the ability to visualize the 5' untranslated region and ribosome binding sites, as well as avoiding a read count inflation problem inherent in short read RNA-seq technologies. More importantly, more biological information could be accessed in long read sequencing of mRNA compared to short read sequencing.

Keywords: RNA sequencing, fragmentation, short read, long read, transcriptome, mRNA transcript, 5' untranslated region, ribosome binding site, next-generation sequencing, third-generation sequencing,

Subject areas: biochemistry, genomics, bioinformatics, molecular biology, cell biology,

Conflicts of interest

The author declares no conflicts of interest.

Funding

No funding was used in this work.