Archives of Science

Cancer Transcriptome-Based Resolution of Isoform Complexity by Pacific Biosciences Fusion and Long Isoform Pipeline

Tyler Jacobs*

Department of Molecular Biology, College of Colchester, United Kingdom

Abstract

Short-read sequencing for genomic profiling is useful for identifying disease-related variation in both DNA and RNA. However, molecular profiling utilising long-read sequencing enhances the resolution of such events because structural variation in cancer occurs often. For instance, the Pacific Biosciences long-read RNA-sequencing (Iso-Seq) transcriptome technique finds expressed fusion partners and offers full-length isoform characterisation, discernment of allelic phasing, and isoform identification. To find expressed fusion partners and isoforms, the Pacific Biosciences Fusion and Long Isoform Pipeline (PB FLIP) uses a variety of RNA-sequencing software analysis tools and scripts. In order to test our methodology and analytical performance, sequencing of a commercial reference (Spike-In RNA Variants) with known isoform complexity was carried out. This sequencing showed strong recall of the Iso-Seq and PB FLIP workflow. This work explains how Iso-Seq and PB FLIP analysis can help with isoform recognition and difficult structural variant deconvolution in a cohort of institutional paediatric and adolescent/young adult cancer research participants. The exemplary case studies show that Iso-Seq and PB FLIP can distinguish between allele-specific expression patterns, resolve complex intragenic changes, and find novel expressed fusion partners.

Keywords: Bioscience; Cancer; Transcriptome

Introduction

A comprehensive perspective of the genetic variation in cancer genomes is now possible because to next-generation sequencing techniques. The range of mutations linked to oncologic illnesses includes structural variation, base alterations, insertion-deletion events, and single-nucleotide variation. The capacity to resolve these genomic changes is significantly influenced by the length of the sequencing reads and the bioinformatic techniques used. From short read lengths of 25 to 35 bp to those utilised today, which range from 100 to 300 bp for current short-read chemistries, next-generation sequencing approaches have progressed. 4 Lengthy-read sequencing platforms, such as single-molecule fluorescence zero-mode waveguide or single-molecule nanopore-based sequencing, are making it more possible to sequence long molecules (>10,000 bp) [1].

Double-stranded DNA or cDNA molecules capped by hairpin adapters (SMRTbell) may now be routinely manufactured up to 15,000 bp in length thanks to Pacific Biosciences' (PacBio) high fidelity and long-read RNA-sequencing (Iso-Seq) methods. 5 The Watson and Crick strands can be sequenced numerous times by the polymerase in an SMRTbell molecule because it is topologically circular and structurally linear. This results in multiple subreads that are separated by the hairpin adapter sequence. The per-base accuracy rate can rival Illumina technology due to the random nature of errors in single-molecule real-time (SMRT) sequencing, which collapses subreads to generate a circular consensus sequence (CCS) [2,3].

Long-read sequencing is uniquely positioned to transform clinical next-generation sequencing applications thanks to its capacity to produce long (5000–15,000 bp range) accurate reads. De novo assembly (as opposed to reference alignment), the characterisation of full-length transcripts, and the resolution of structurally difficult genomic areas are all benefits of long-read sequencing. Additionally, Iso-Seq eliminates the need for RNA fragmentation and avoids the intrinsic constraint of short-read RNA sequencing by maintaining the expressed exonic order and orientation (RNA-Seq). 9 As a result, Iso-Seq represents individual transcripts that potentially reveal brand-new isoforms linked to disease. For instance, Iso-Seq has proven to be clinically useful in locating novel

variant oncogene isoforms in gastric cancer cell lines [4].

Materials and Method

Sample Preparation and Subsequent Sequencing for PacBio Iso-Seq

300 ng of total tumour or HBR RNA spiked with a 2% SIRV-Set 4 synthetic RNA mix were utilised for first-strand cDNA synthesis (Spike-In RNA Variants; Lexogen, Vienna, Austria; catalogue number 141). For the remainder of the text, the HBR RNA sample spiked with 2% SIRV-Set 4 will be referred to as HBR/SIRV. The response adhered to the Single Cell/Low Input NEBNext methodology. Iso-Seq Express Template Procedure and Checklist's oligo (dT) priming of polyadenylated mRNA is used in this technique. In the Illumina (San Diego, CA) methodology, the ribodepletion step is eliminated by iso-Seq [see below; PN 101 to 763-800 version 02 (October 2019)]. Size selection of the resultant cDNA at >1000 bp was carried out using ProNex Beads were used to size select the resultant cDNA at >1000 bp, and the final elution volume was 17 L in Buffer EB (Qiagen, Hilden, Germany). Takara PrimeStar GXL DNA polymerase was then utilised for further PCR amplification (2.5 U; Takara, Shiga, Japan). The PCR primer combination contained one primer from the PacBio Iso-Seq Express kit and one primer from the NEBNext Single Cell kit (catalogue number E6421S) (PN number 101 to 737-500; PacBio, Menlo Park, CA). The cDNA was amplified by PCR using two 50 L PCRs each containing eight litres of purified first-strand cDNA

***Corresponding author:** Tyler Jacobs, Department of Molecular Biology, College of Colchester, United Kingdom, E-mail: TylerJ33@gmail.com

Citation: Jacobs T (2022) Cancer Transcriptome-Based Resolution of Isoform Complexity by Pacific Biosciences Fusion and Long Isoform Pipeline. Arch Sci 6: 139.

Citation: Jacobs T (2022) Cancer Transcriptome-Based Resolution of Isoform Complexity by Pacific Biosciences Fusion and Long Isoform Pipeline. Arch Sci 6: 139.

Page 2 of 2

template, one litre of PrimeSTAR GXL buffer, 0.1 mmol/L of dNTPs, 1.25 units of PrimeSTAR GXL DNA polymerase (1.25 U/L; Takara Bio, San Jose, CA), one litre of NEBNext single cell, and one litre of ISO-Seq Express The following PCR cycling parameters were used: 30 s at 98°C, 16 cycles (10 s at 98°C, 15 s at 65°C, and 10 m at 68°C), and 5 m at 68°C [5,6,7].

Preparation and Sequencing of HBR/SIRV Short-Read Illumina RNA-Seq Libraries

The mRNA content for HBR is listed in the Lexogen Spike-In RNA Variant control user guide (SIRV-Set 4; Lexogen catalogue number 141) as being 2% of total HBR RNA. Therefore, 500 ng total HBR RNA was used as the starting RNA material for library construction, and 200 pg of SIRV-Set 4 was added to achieve a 2% SIRV spike-in within the final library.

Ribodepletion is necessary for Illumina RNA-Seq libraries produced using total RNA. NEBNext rRNA Depletion (NEB number E6310), RNA fragmentation (5 minutes), and cDNA conversion were all steps in the processing of the HBR/SIRV sample [8, 9].

Processor of PacBio SMRT Link Data

The web-based PacBio SMRT Link version 10.0.0.108,728 software is used to examine iso-Seq data, and it includes apps for designing sequencing runs, managing data, and assisting with secondary data analysis [10].

Discussion

Patients who sign up for our IRB protocol have their exomes analysed in pairs using samples from the disease and a germline control. The variation in the protein-coding areas of the genome can be studied thanks to this profiling. It gives information on the genetic diversity that contributes to disease, potentially improving the diagnosis and prognosis of rare and refractory cancers and haematological disorders. Moreover, the genetic landscape in juvenile malignancies is unique compared with adult tumours, with an overall lower mutational burden. Other chromosomal anomalies, like structural variants, can help to partially explain the genesis of paediatric tumour formation. Even yet, resolving intragenic insertions, deletions, and gene fusions with short-read exome data is frequently challenging. Long-read genome sequencing can be used in situations when paired-exome sequencing is ineffective at producing a diagnostic result because of chromosomal abnormalities.

Conclusion

In conclusion, the processes for iso-Seq and PB FLIP outlined here enable the ongoing processing of N-of-one clinical samples while resolving complex somatic changes. The requirement for high-molecular-weight and high-quality nucleic acids to be separated from samples, however, poses a hurdle to the use of long reads. Clinical samples are typically preserved using formalin fixation and paraffin embedding, which causes nucleic acids to become cross-linked, damaged, and degraded58,59. It is therefore challenging to get detailed long-read information from formalin-fixed, paraffin-embedded samples. The relative cost per sample of using Iso-Seq for clinical sequencing presents another difficulty. This cost is higher than that of RNA-Seq. Utilizing kilobase read durations is one method for lowering costs. We describe our capability to accurately characterise and sequence each of the lengthy SIRV transcripts up to 12,000 base pairs. The number of full-length cDNAs per SMRTbell molecule can be increased via cDNA concatenation thanks to these lengthy read lengths. 60 Iso-Seq offers a new area of excitement focused on alternative splicing in healthy and disease states to better assess the unique isoforms that may regulate phenotypic differences thanks to advancing technologies and related cost reductions.

Potential Conflicts of Interest

The author has no conflict of interest.

References

1. Ferraro, NM (2020) Transcriptomic signatures across human tissues identify functional rare genetic variation. Science369.

2. Wang ET (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456: 470-476.

3. Djebali S (20120 Landscape of transcription in human cells. Nature 489: 101-108.

4. Lendahl U, Lee KL, Yang H, Poellinger L (2009) Generating specificity and diversity in the transcriptional response to hypoxia. Nat Rev Genet 10: 821-832.

5. Monticelli S, Natoli G (2017) Transcriptional determination and functional specificity of myeloid cells: making sense of diversity. Nat Rev Immunol 17: 595-607.

6. Xiang Y, Ye Y, Zhang Z, Han L (2018) Maximizing the utility of cancer transcriptomic data. Trends Cancer 4: 823-837.

7. Wu J (2021) Maximizing the utility of transcriptomics data in inflammatory skin diseases. Front Immunol 12: 761890.

8. Kahles A (2018) Comprehensive analysis of alternative splicing across tumors from 8,705 patients. Cancer Cell 34: 211-224.

9. Xiang Y (2018) Comprehensive characterization of alternative polyadenylation in human cancer. J Natl Cancer Inst 110: 379-389.

10. Guo W (2018) A LIN28B tumor-specific transcript in cancer. Cell Rep 22: 2016-2025.