Long read transcriptomes identify features not found with very deep short read sequencing

Sykes, M., Pionzio, A.M., Hildebrand, S., Singh, S., Lakatos, R., Prasad, N., Umylny, B, Discovery Life Sciences

DISCOVERY LIFE SCIENCES

Science at your Service.™

INTRODUCTION

The limited capacity of long-read sequencing technologies has significantly limited researchers' ability to efficiently apply long read sequencing data in drug development. Discovery Life Sciences (DLS) has addressed this challenge by building a service laboratory that utilizes over 20 Sequel® IIe devices from Pacific Biosciences (PacBio) and incorporates liquid handling and robotics to optimize efficiency, cost, and accuracy of large-scale, long-read sequencing projects. To demonstrate the ability of this offering to discover novel isoforms that cannot be detected using established short read RNA-Seq technologies, DLS partnered with researchers to analyze 12 human tumor samples using between 1 and 5 SMRT Cells per sample (3-18 million reads). When compared to short read sequencing results of the same samples, the data shows that even with only 3 million reads, long read technology can identify validated isoforms missed by very deep short read sequencing. Analysis by quality control and annotation tools further shows that not only are the isoforms identified by short read sequencing much shorter, but they also are much more likely to be false positives, making analysis of the short read transcriptomes more difficult.

METHODS & MATERIALS

Ribosomal RNA (rRNA) reduction RNA-seq library preparation and sequencing

The concentration and integrity of the RNA was estimated by Quant-it™ Ribogreen assay (Thermo Fisher), and Fragment Analyzer (Agilent), respectively. Approximately 500ng of total RNA from each sample was taken into library prep using the Illumina Stranded Total RNA Prep with Ribo-Zero Plus kit (Illumina) per manufacturer's recommended protocol. Final library concentration was measured by Quant-it™ Picogreen™ Assay (Thermo Fisher), and the library size was estimated by utilizing a DNA High Sense chip on LabChip® Gx Touch™ analyzer (PerkinElmer). Accurate quantification of the final libraries for sequencing applications was determined using the qPCR-based KAPA® Biosystems Library Quantification kit (Roche). 2x100 PE Sequencing was performed on a NovaSeq® 6000 instrument (Illumina).

Iso-Seq® library preparation using SMRTbell® prep kit 2.0

The concentration and integrity of the RNA was estimated by Quant-it™ Ribogreen assay (Thermo Fisher), and Fragment Analyzer (Agilent), respectively. Approximately 300ng of total RNA from each sample was taken into cDNA generation using the NEBNext® Single Cell/Low Input cDNA Synthesis & Amplification kit. Post cDNA libraries were generated using SMRTbell® Express Template Prep kit 2.0 per manufacturer's recommended protocol. Each library was sequenced on an individual SMRTcell on Sequel® IIe instrument (PacBio).

Data Analysis

For each sample, a transcriptome was generated from Illumina short read data (2X150 bp) using the Trinity assembler (1) with the Trimmomatic (2) option enabled. These transcriptomes were mapped to human genome build hg38 using minimap2 (3), and then collapsed using cDNA cupcake (4). A second transcriptome for the same sample was generated from PacBio long read data using the IsoSeq pipeline from PacBio. Both short read and long read transcriptomes were annotated using SQANTI3 (5).

In addition to the SQANTI3 QC metrics, a database of signature cancer splice junctions was used to assess transcriptome quality. To determine if these junctions were detected in a transcriptome, the junction sequences were aligned to the transcriptome using bwa mem (6). A junction was considered present if the entire junction mapped to the transcriptome without insertions, deletions or clipping. Mismatches were acceptable if they did not prevent mapping using default bwa mem alignment parameters.

RESULTS

While Illumina short-read sequencing is still the most cost effective and efficient method for generating large quantities of genomic and transcriptomic data, improved accuracy and software of the PacBio sequencers offers new opportunities for discovering full length novel transcripts.

Isoform Classification

A significantly higher percentage of the PacBio transcripts represent full-length isoforms, while most Illumina/Trinity isoforms appear to come from intergenic regions (Figure 1), even though intergenic reads represent a small fraction of the total sequencing (many transcripts sequenced very few times each).

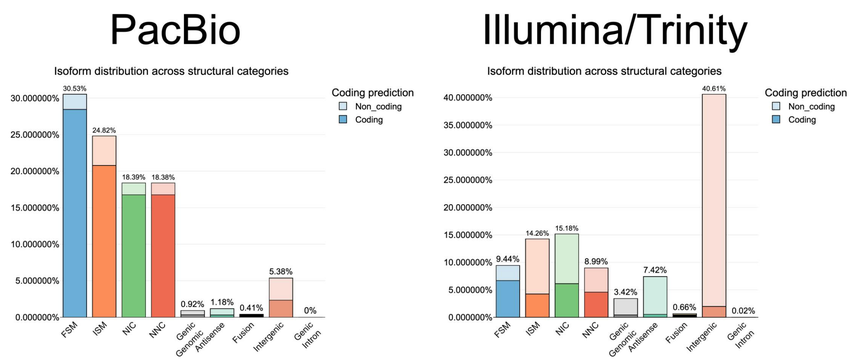

Figure 1: Isoform classification

Table 1: SQUANTI3 Legend

Acronym	Description
FSM	Matches all splice junctions perfectly
ISM	Matches the reference splice junctions partially
NIC	Novel isoform with a new combination of known splice sites
NNC	Novel isoform with at least 1 new splice site

Cage Distances and PolyA Distances Analyses

Significantly higher percentage of PacBio isoforms are within close proximity of CAGE (Figure 2) and PolyA (Figure 3) peaks.

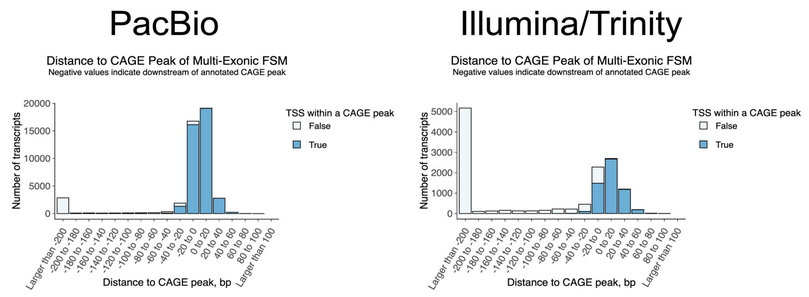

Figure 2: Distance from PacBio (left) and Illumina (right) FSM isoforms to CAGE peaks

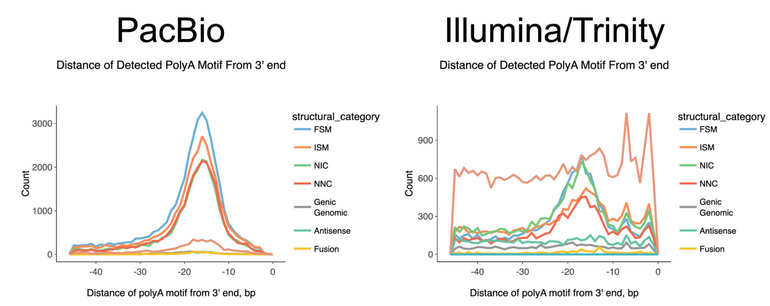

Figure 3: Distance from PacBio (left) and Illumina (right) isoforms to PolyA

Transcript Length

PacBio resulted in significantly higher proportion of longer transcripts (Figure 4)

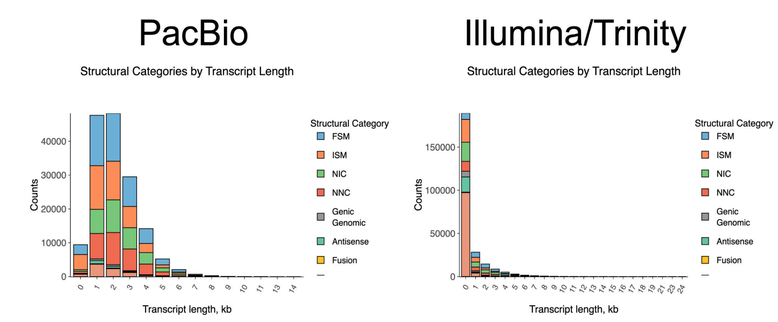

Figure 4: Isoform length distribution for PacBio (left) and Illumina (right)

Total Genes, Annotated Genes, and Annotated Transcripts, and Isoforms

Illumina does detect many more genes, both total (Figure 5) and novel (Figure 6). However, as the figures show, PacBio data are not yet saturated and additional sequencing can be expected to identify more genes.

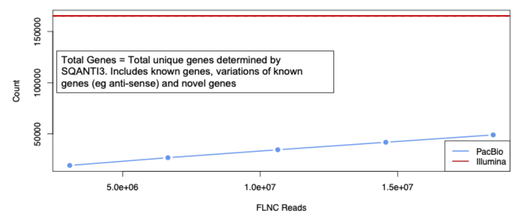

Figure 5: Total genes identified by PacBio (blue) and Illumina (red)

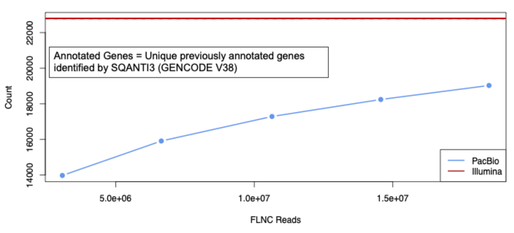

Figure 6: Annotated genes identified by PacBio (blue) and Illumina (red)

PacBio, given sufficient sequencing, does identify more annotated transcripts (Figure 7) and unique isoforms (Figure 8).

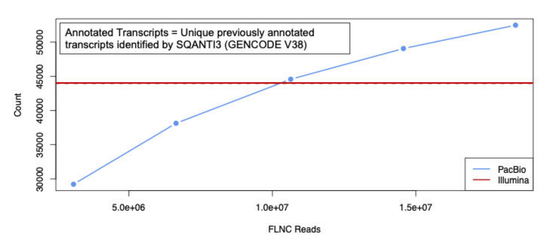

Figure 7: Annotated transcripts identified by PacBio (blue) and Illumina (red)

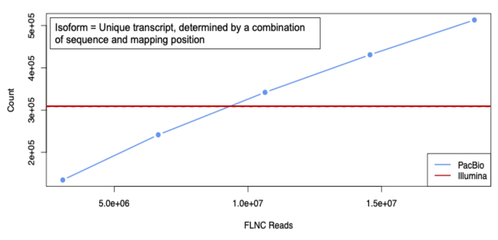

Figure 8: Unique isoforms detected by PacBio (blue) and Illumina (red)

Identification of Splice Junctions

Searching specifically for known splice junctions reveals a different set of junctions identified by Illumina and PacBio platforms (Table 2). As expected, increased PacBio sequencing allows identification of greater number of junctions.

Table 2: Number of junctions identified by Illumina (column 3) and PacBio (columns 4-8) based on the number of PacBio SMRTcells (1-5). For PacBio, numbers in brackets indicate the number of junctions also identified by Illumina.

Sample	Cancer	Illumina	PacBio 1	PacBio 2	PacBio 3	PacBio 4	PacBio 5
0001	other	36	5 *(4)*				
0002	other	44	6 *(4)*				
0003	other	54	18 *(13)*	26 *(16)*			
0004	other	45	8 *(6)*				
0005	other	38	9 *(6)*				
0006	other	30	6 *(3)*				
0007	other	39	10 *(8)*	15 *(11)*			
0008	signature	63	14 *(9)*	24 *(16)*	30 *(17)*		
0009	signature	54	17 *(12)*	27 *(16)*			
0010	signature	37	6 *(2)*	10 *(6)*	18 *(8)*		
0011	signature	59	16 *(13)*	32 *(22)*	39 *(26)*	44 *(27)*	49 *(29)*
0012	signature	37	17 *(6)*	24 *(11)*			

CONCLUSIONS

- While larger data sets allow Illumina to identify broad set of features, longer reads allow PacBio sequencing to identify features missed using Illumina-only approach. It also should be noted that even at 5 SMRT cells the PacBio data set is not yet saturated, so we can expect to identify more features with additional sequencing.
- PacBio approach results in identification of different features than Illumina (Table 2), leading to a possibility that an approach utilizing both short-reads and long-reads sequencing may be advantageous.