Full-length Isoform Sequencing of the Human MCF-7 Cell Line using PacBio® Long Reads

Elizabeth Tseng[1], Tyson Clark[1], Meredith Ashby[1], Gloria Sheynkman[2]
[1]Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025
[2]Center for Cancer Systems Biology, Dana Farber Cancer Institute, and Department of Genetics, Harvard Medical School

Introduction

While advances in RNA sequencing methods have accelerated our understanding of the human transcriptome, isoform discovery remains a challenge because short read lengths require complicated assembly algorithms to infer the contiguity of full-length transcripts. With PacBio's long reads, one can now sequence full-length transcript isoforms up to 10 kb. The PacBio Iso-Seq™ protocol produces reads that originate from independent observations of single molecules, meaning no assembly is needed.

Here, we sequenced the transcriptome of the human MCF-7 breast cancer cell line using the Clontech SMARTer® cDNA preparation kit and the PacBio RS II. Using PacBio Iso-Seq bioinformatics software, we obtained 55,770 unique, full-length, high-quality transcript sequences that were subsequently mapped back to the human genome with ≥ 99% accuracy. In addition, we identified both known and novel fusion transcripts. To assess our results, we compared the predicted ORFs from the PacBio data against a published mass spectrometry dataset from the same cell line. 84% of the proteins identified with the Uniprot protein database were recovered by the PacBio predictions. Notably, 251 peptides solely matched to the PacBio generated ORFs and were entirely novel, including abundant cases of single amino acid polymorphisms, cassette exon splicing and potential alternative protein coding frames.

Iso-Seq™ Library Preparation & Bioinformatics Workflows

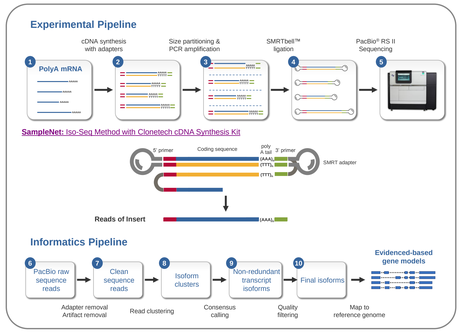


Figure 1: The Iso-Seq Method sample prep and analysis workflow.

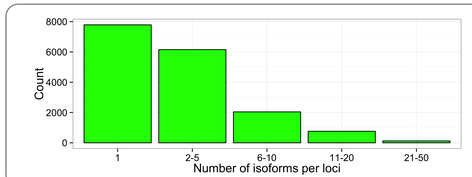
The Iso-Seq Method can reveal new information about:

- Alternative splicing
- Alternative polyadenylation
- Novel genes
- Non-coding RNAs
- Fusion transcripts

Table 1. Sample prep and data collection.

Size Selection Method	Size Binning	Sequencing Chemistry	Total SMRT® Cells
Agarose Gel	1-2 kb, 2-3 kb, 3-6 kb	P4-C2	119
SageELF™ System	1-2 kb, 2-3 kb, 3-5 kb, 5-10 kb	P6-C4	28

Full-Length Isoform Characterization of MCF-7



Total genomic loci: 16,868
Non-redundant isoforms: 55,770

Min transcript length: 315 bp
Max transcript length: 9,436 bp
Mean transcript length: 2,134 bp

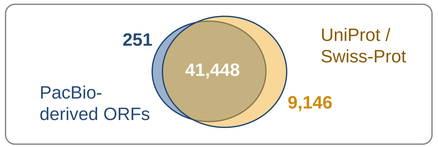
Figure 2. Number of isoforms per locus. Transcripts that overlap on the genomic coordinate by 1 bp on the same strand are grouped together to form non-overlapping transcribed loci. Out of 16,868 expressed loci, 149 had more than 20 isoforms, with a total average of 3 isoforms per loci.

Mass Spectrometry Validation

	Total (PB+ UniProt)	PacBio ORFs	UniProt / Swiss-Prot
peptides	50,845	41,699	50,594
proteins*	n/a	6,718	8,010

Table 2. MCF-7 mass spec data[1] was validated against PacBio and Uniprot protein databases. *Proteins are grouped by genetic locus.

Figure 3. PacBio sequencing-derived ORFs matched 251 novel peptide not found in Uniprot. Of the 9,146 missed peptides (matched by Uniprot only), >95% were long transcripts, and <1% were genes with poly(A)-transcripts.



Variation	Frequency
Single amino acid polymorphism	158
Non-canonical start site	33
Alternative splice	22
Frameshift	12
'Non-coding' RNA, short ORF	11
UTR	5
Novel gene	5
Insertion	3
Deletion	2

Table 3. The PacBio ORF database enabled the identification of novel peptides from the mass spec data that were not present in Uniprot, including non-canonical start sites, alternative splice events, and novel genes, as well as evidence of cell line specific genomic structural variations such as single amino acid polymorphisms (SAP), frameshifts, insertions and deletions.

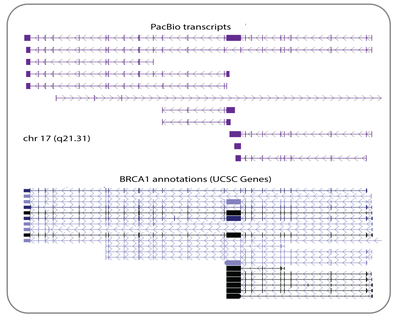
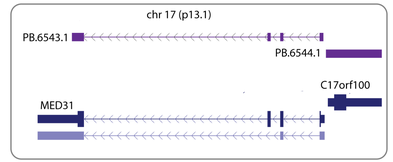


Figure 4. (left) PacBio transcripts capture multiple isoforms of the BRCA1 gene, several of which are novel.

Figure 5. (below) Antisense gene pair KIAA0753-MED31. This is a known gene pair that has been previously reported.



Fusion Transcript Discovery

Gene1	Chrom1	Gene2	Chrom2
BCAS3	chr17	BCAS4	chr20
EIF3D	chr22	MYH9	chr22
RPS6KB1	chr17	DIAPH3	chr13
SYTL2	chr11	PICALM	chr11
SYTL2	chr11	VMP1	chr17
SULF2	chr20	ARFGEF2	chr20
SULF2	chr20	ZNF217	chr20
FOXA1	chr14	TTC6	chr14

Table 4. A total of 104 fusion candidates were identified, including multiple fusion points for many known gene fusion pairs. Enumerated here is a select list of PacBio-identified fusion transcripts supported by previous literature.

Conclusions

The Iso-Seq method provides an opportunity for researchers to make new discoveries about the complex splicing events that occur in cancer cells. Unlike short-read alternatives, PacBio long-read sequencing of full-length transcripts allows for the discovery of isoforms from 300 bp to 10 kb with no assembly required, enabling the discovery of novel isoforms and transcribed gene fusions.

References and Acknowledgments

1. Geiger, et. al. (2012) Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol Cell Proteomics*. **11**(3):M11.014050.
2. Link to UCSC Genome browser track for this data set

The authors would like to thank everyone who helped generate data for this poster.