Using the PacBio Iso-Seq method to search for novel colorectal cancer biomarkers

Meredith Ashby[1], Malgorzata A. Komor[2], Annemieke C. Hiemstra[2], Thang V. Pham[3], Sander R. Piersma[3], Robert P. Sebra[4], Bo W. Han[1], Beatriz Carvalho[2], Gerrit A. Meijer[2], Connie R. Jimenez[3] and Remond JA Fijneman[2]

[1]PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025; [2]Netherlands Cancer Institute, Amsterdam, NL; [3]VU University Medical Center, Amsterdam, NL; [4]School of Medicine at Mount Sinai, Institute for Genomics and Multiscale Biology, New York, NY

Introduction

Early detection of colorectal cancer (CRC) and its precursor lesions (adenomas) is crucial to reduce mortality rates. The fecal immunochemical test (FIT) is a non-invasive CRC screening test that detects the blood-derived protein hemoglobin. However, FIT sensitivity is suboptimal (~65%) especially in detection of CRC precursor lesions (~27%). New biomarkers are needed to improve early detection and increase cure rates.

As adenoma-to-carcinoma progression is accompanied by changes in mRNA splicing, tumor-specific proteins derived from alternatively spliced RNA transcripts might serve as candidate biomarkers for CRC detection. To investigate this hypothesis, existing annotation databases were supplemented with new, highly pertinent isoform information by sequencing mRNA and proteins from a CRC cell line before and after down-modulation of splicing machinery using both NGS and full-length isoform sequencing from PacBio.

Using this approach, numerous candidate proteins were identified for follow-up evaluation in patient samples. In addition, the Iso-Seq method from PacBio was shown to be more effective than Illumina RNAseq in uncovering several categories of alternative splicing events.

Challenge of Incomplete Databases

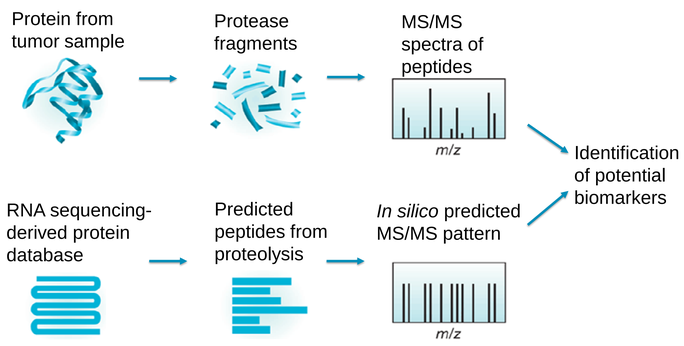


Figure 1. Using MS/MS to identify novel biomarkers. Identification of proteins depends on associating observed spectra with the predicted spectra of known proteins, often derived from RNA sequencing databases. A key challenge is that ~50% of mass spectra are commonly not identifiable. Augmenting databases with full-length isoform sequencing of underrepresented classes of proteins may improve our ability to uncover previously hidden biomarkers.

Experimental Design

Both mRNA and proteins were isolated from CRC cell line SW480 before and after siRNA-mediated down-modulation of the splicing machinery SF3B1. To identify splice variants, SW480 siSF3B1 and control samples were analyzed for alternatively spliced mRNA transcripts using the PacBio Iso-Seq method, which generates full-length mRNA sequences, and with RNAseq using Illumina.

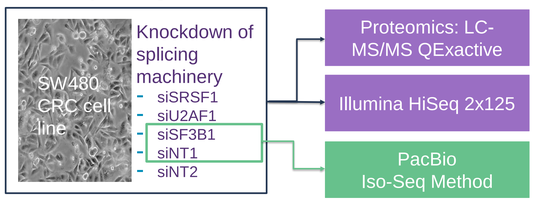


Figure 2. Workflow for developing a protein annotation database enriched for cancer-relevant abnormal splicing events.

PacBio Sequencing Reveals Candidate Biomarkers

The data revealed hundreds of mRNA splice variants, including positive controls described in literature. For example, down-modulation of SF3B1 resulted in shifts among spliced isoforms of *ADD3*.

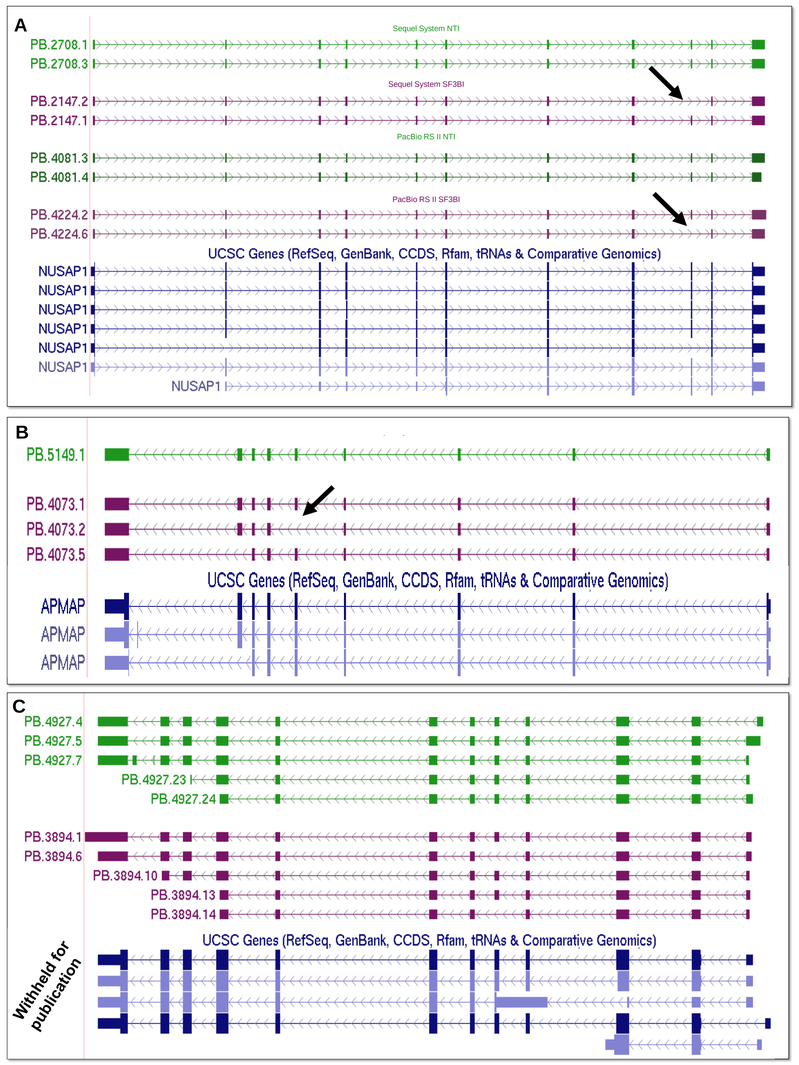
Table 1. Summary of Iso-Seq data generated on the PacBio RS II platform.

SAMPLE	INSERT SIZE	# OF ZMWS	# OF SUBREADS	# OF CCS	# OF FL	FL%
NTI	1 - 2 kb	102,506	**750,417**	77,851	24,268	31%
NTI	1 - 2 kb	80,952	**523,721**	66,802	13,174	20%
NTI	2 – 3 kb	115,990	**660,651**	96,273	41,326	43%
NTI	2 – 3 kb	111,296	**658,674**	98,348	42,551	43%
SF3BI	1 – 2 kb	63,855	**321,124**	47,511	5,543	12%
SF3BI	1 – 2 kb	111,585	**981,722**	86,846	42,632	49%
SF3BI	2 – 3 kb	113,143	**797,887**	97,332	51,726	53%
SF3BI	2 – 3 kb	105,585	**620,021**	87,917	38,482	44%

Table 2. Summary of Iso-Seq data generated on the Sequel platform.

SAMPLE	INSERT SIZE	# OF ZMWS	# OF SUBREADS	# OF CCS	# OF FL	FL%
NTI	1 - 2 kb	568,279	**11,304,769**	144,547	52,857	37%
NTI	2 - 3 kb	682,189	**1,847,492**	177,435	90,902	51%
SF3BI	1 - 2 kb	756,935	**8,658,168**	96,366	35,085	36%
SF3BI	2 - 3 kb	851,135	**1,712,813**	125,919	63,313	50%

Figure 3. Examples of isoforms found in Iso-Seq data. Several examples of isoforms found in control NT1 (green) and splicing compromised 'SF3B1' (purple) samples. **(A)**. Isoforms are reproducibly seen on both the PacBio RS II and Sequel platforms. **(B)**. PacBio data reveals novel isoforms in splicing-compromised CRC cells. **(C)**. A number of the genes found to have novel isoforms are pertinent to colorectal cancer development and will be validated in patient samples.



PacBio Iso-Seq Method Excels at Isoform Discovery

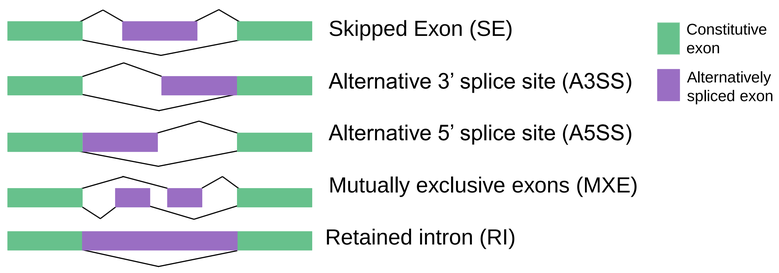
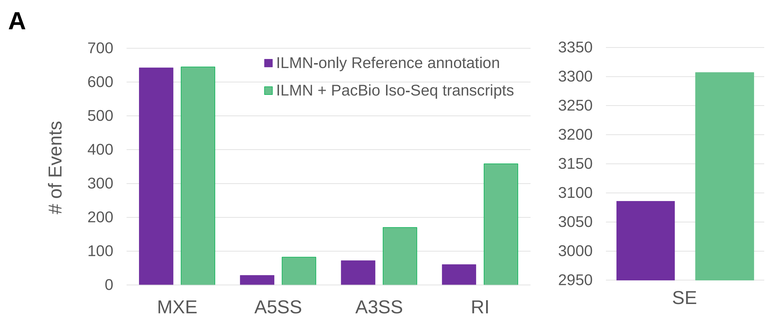


Figure 4. Types of alternative splice events found in perturbed SW480 cell lines.



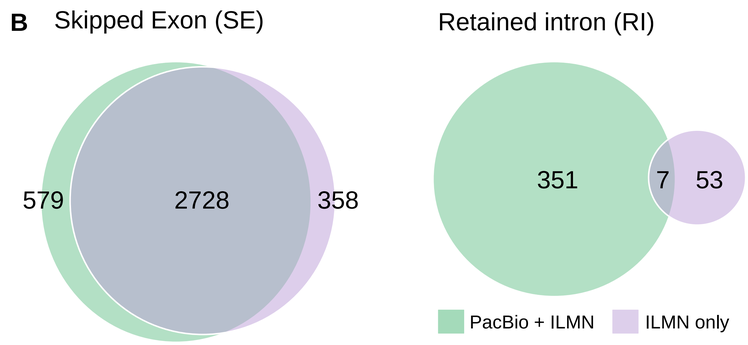


Figure 5. Efficiency of alternative splice discovery. (A) The above chart shows how many of each type alternative splicing event was detected by combining short read and Iso-Seq data (green) versus RNAseq only (purple) in the SF3B1 knockdown sample, highlighting the advantage of full-length isoform sequencing for detecting several categories of alternative splicing events. **(B)** Venn diagrams show that even at relatively low coverage, adding PacBio sequencing results in detection of many events missed by short reads alone.

Conclusions

- PacBio long-read sequencing reveals full-length isoforms with no ambiguity, bringing clarity to complex biological systems.
- Adding Iso-Seq data revealed more skipped exons, alternative 3' and 5' splice sites, and retained introns in a splicing-compromised model system as compared to NGS sequencing alone.
- We are currently applying the mRNA sequencing based proteogenomic pipeline for detection of protein alterations to a series of adenomas at low- and high-risk of progression, and CRC tumor samples. Novel findings will be evaluated for their performance as screening markers for CRC.
- The unique strengths of PacBio long-read sequencing make it a powerful tool for biomarker discovery.