Improving isoform detection sensitivity for PacBio Iso-Seq using CRISPRclean® Single Cell RNA Boost kit

Depletion increases opportunity for sequencing of low expression genes and isoforms

Summary

Data collected in this study with PBMCs demonstrates that using the CRISPRclean Single Cell RNA Boost Kit to deplete noisy uninformative molecules in a sample prior to sequencing results in accurately detecting more unique molecules, genes, and isoforms than increasing sequencing depth alone can attain. Results from samples depleted with the CRISPRclean Single Cell RNA Boost Kit contain 7.23-fold more unique molecules, 1.9-fold more unique genes, and 4.5-fold more unique isoforms compared to the non-depleted samples at the same sequencing depth. Subsampled data sets show that increased sequencing depth is limited in new isoform detection and cannot match the results from the depleted set. This demonstrates the large boost in isoform detection achieved with depletion cannot be matched by simply increasing sequencing depth of a non-depleted sample.

Introduction

Single cell RNA sequencing (scRNA-seq) technologies have helped revolutionize transcriptome studies due to its unique ability to characterize the transcriptome of a cell on an individual level with unprecedented detail.[1,2] Alternative splicing events can be discovered with both short and long read systems.[1-3] However, to scale up this method and increase user adoption, the cost of sequencing and ability to detect low expression genes of interest need to be improved. Additionally, short read approaches do not span successive exon boundaries, limiting the ability to detect novel isoforms. Here, we present an approach based on the CRISPR-Cas9 system, marketed as CRISPRclean, that removes uninformative fragments from scRNA-seq libraries prior to sequencing (Figure 1) for long read sequencing technology from Pacific Biosciences (ISO-Seq). By doing so, we greatly increase informative reads in scRNA-seq data and enable the discovery of additional isoforms and cell types.
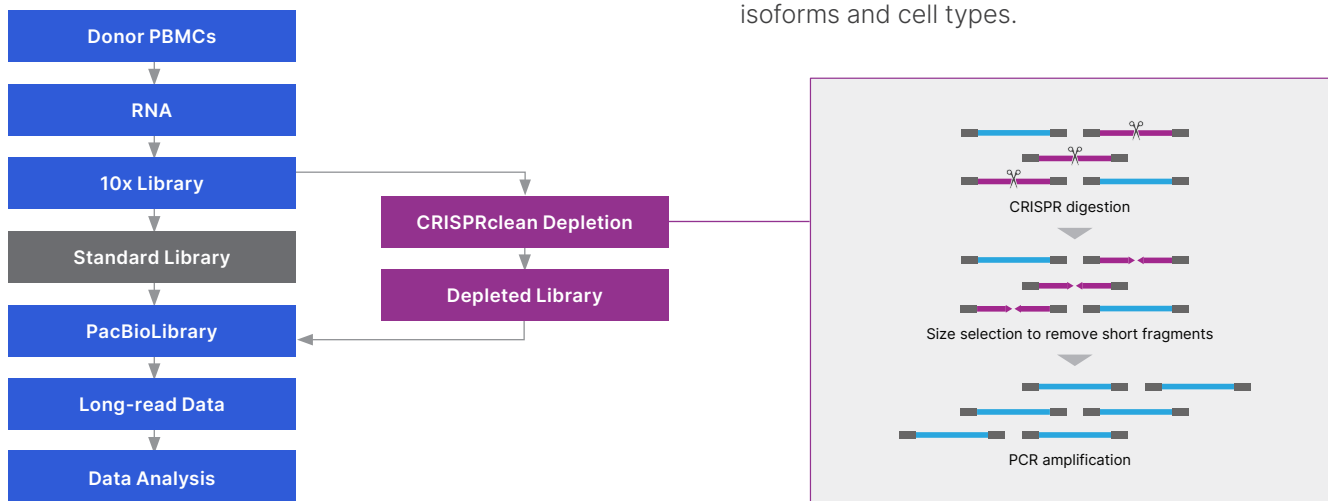

Figure 1: Overview of CRISPRclean technology. Workflow showing samples prepared for Pacific Biosciences (ISO-Seq) both with CRISPRclean depletion (purple) without depletion (grey). Both sets of samples were sequenced to the same sequencing depth and processed for secondary analysis.
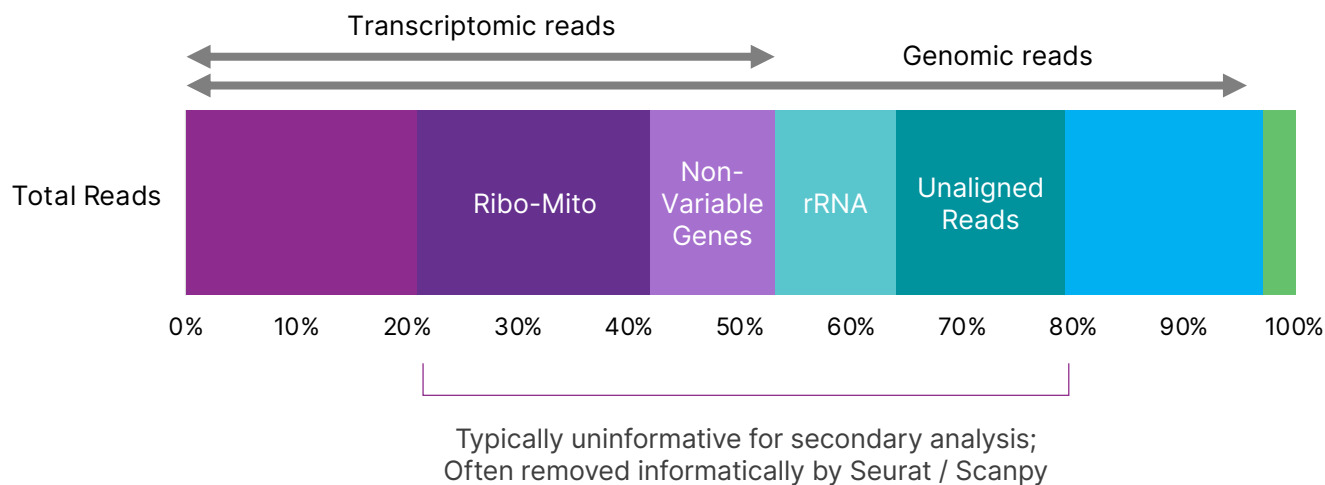
Figure 2: Visual Representation of targeted content for depletion in single cell boost (SCB). Results represent typical read distribution from a single cell sequencing run. Four regions for removal are highlighted. Ribosomal and mitochondrial (Ribo-Mito) genes and non-variable genes from the transcriptomic reads are the first two regions highlighted. Ribosomal RNA (rRNA) and unaligned reads from genomic reads also targeted for removal.

Methods

RNA was extracted from donor PBMC samples and processed using the Chromium Next GEM Single Cell 3' v3.1 dual index kit from 10x Genomics. Single cell Iso-Seq libraries were then prepared using the PacBio SMRTbell Express Template Prep Kit 2.0 with and without CRISPRclean depletion. CRISPR guide RNAs were designed against genes and genomic intervals identified for removal from sequencing libraries. These guides were targeted to the single cell boost targets (SCB), described in Figure 2.

Unique guide RNAs were synthesized for testing in the CRISPRclean method. The resulting libraries were sequenced on the PacBio Sequel II sequencing platform using 1 SMRT Cell 8M per sample. The resulting data from the control and depleted samples was analysed in 4 ways:

1. Percentage of reads aligning to each targeted region
2. Depletion rates of targeted molecules
3. Total isoforms
4. Novel isoforms

Results

Comparison of the results of the depleted condition to the control condition shows a significant reduction in coverage of the targeted content, boosting the signal to noise in the sample. Figure 3 shows that the SCB

guide set efficiently removes targeted molecules from single cell Iso-seq libraries. For the control sample, targeted content for removal occupies nearly 25% of the aligned reads. When removing that content in the depleted condition, reads are returned to other regions of the transcriptome, allowing increased opportunity for sequencing of low expression genes.
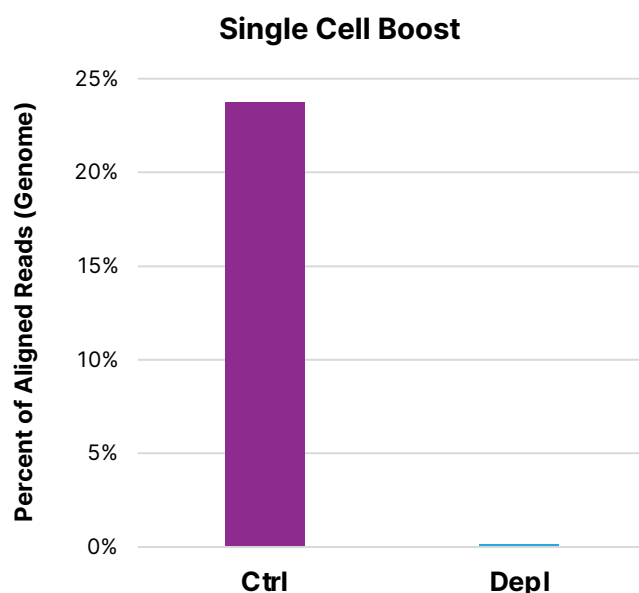

Figure 3: Depletion shows reduced read alignments for targeted regions. The percent of reads mapping to targeted genes and intervals (y-axis) was computed for control (CTRL, purple) and depleted (DEPL, blue) samples. Reads from targeted regions are almost completely removed by the CRISPRclean method.
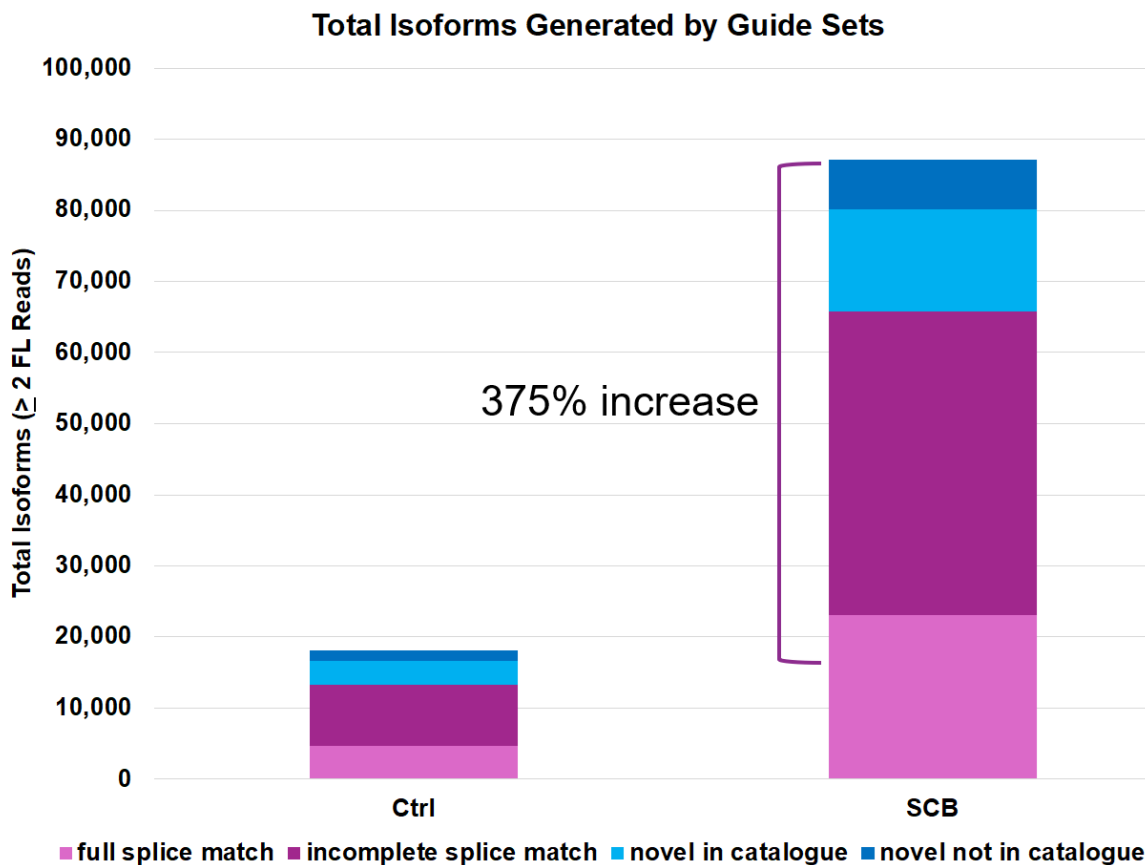
Total Isoforms Generated by Guide Sets

Figure 4: More isoforms are identified in CRISPRclean treated libraries than in untreated conditions at the same sequencing depth. Total and novel isoform species were identified and counted in PacBio Iso-seq data and compared across control and depleted samples. Total isoforms species with at least two full length reads (FL Reads, y-axis) were calculated for control (CTRL) and depleted (DEPL) samples. This guide set enabled the discovery of more isoforms in depleted samples: an increase of 375%.

Additionally, to assess the effect of CRISPRclean depletion on isoform discovery, samples at the same sequencing depth for both the control and depleted conditions were compared. A full splice match is defined by a transcript which matches a reference transcript at all splice junctions while an incomplete splice match is defined as a transcript that matches consecutive, but not all, splice junctions of the reference transcript. Novel in catalogue (NIC) transcripts contain new combinations of already annotated splice junctions or novel splice junctions formed from already annotated donors and acceptors. Novel Not in Catalogue (NNC) transcripts use novel donors and/or acceptors.[6] Figure 4 shows that the CRISPRclean method enables the discovery of more total and novel isoforms in depletions using the single cell boost (SCB) panel. This is possible because CRISPRclean depletion removes fragments prior to sequencing and enables more coverage of informative and lower expressed molecules.

Finally, samples were deduplicated based on UMIs. The non-depleted condition collapsed to 160K reads while the depleted condition collapsed to 1M reads where reads correspond to unique molecules. The samples were then bucketed into 10,000 read bins (Figure 5) to compare filtered reads from the control (purple) to the depleted condition (blue). The number of unique genes (panel A) and isoforms (panel B) was calculated for each bin (y-axis). Data from the depleted sample contains 7.23-fold more unique molecules (panel B) and 1.9-fold more unique genes (panel A). Additionally, the data shows 4.5-fold more unique isoforms compared to the control (data not shown). This data demonstrates the large boost in isoform sensitivity achieved with depletion, a level that cannot be matched by simply increasing sequencing depth of a non-depleted sample.
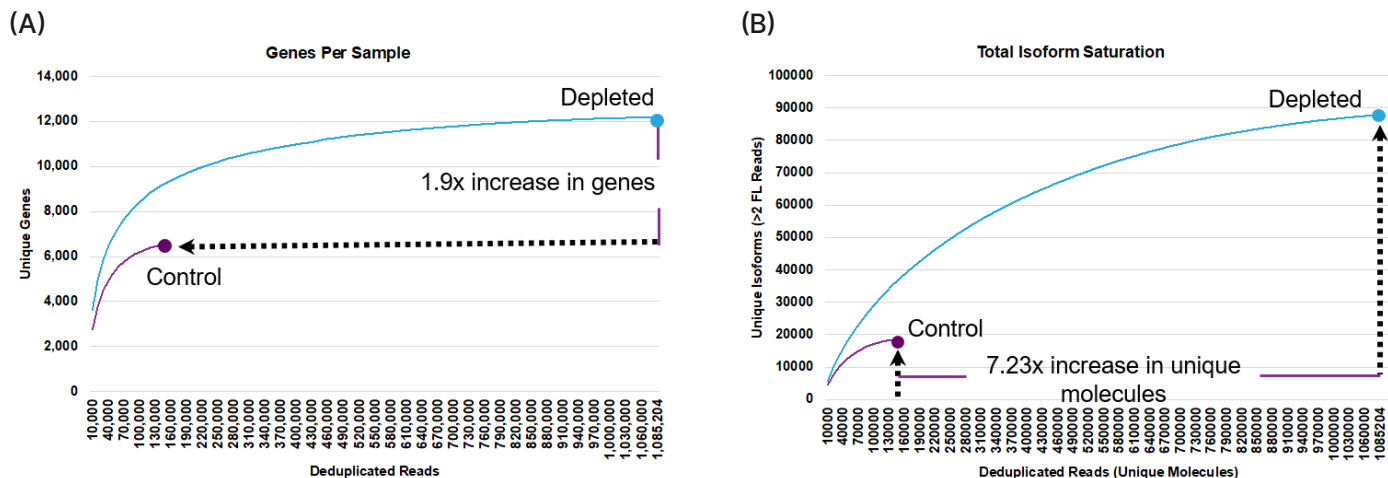
(A)

(B)

Figure 5: Depletion increases information gained and unique molecules in single cell IsoSeq samples. Filtered reads from control (purple) and depleted (blue) PBMC samples were down sampled in 10,000 read bins (x-axis) and the number of unique genes (panel A) and isoforms (panel B) was calculated for each bin (y-axis). Data from the depleted sample contained 7.23-fold more unique molecules, 1.9-fold more unique genes and 4.5-fold more unique isoforms than the control. Further, this data demonstrates that more sequencing would not enable increased discovery in the control sample and that only depletion provides the ability to increase discovery metrics.

Conclusions

Single cell transcriptome studies, while offering unprecedented characterization of transcript expression at the single cell level, suffer from decreased sensitivity since data from highly expressed genes can obscure signals from lower expressed genes. The CRISPRclean method helps mitigate this effect by removing fragments produced by highly expressed genes, prior to sequencing, thus preserving sequencing space for informative molecules. The results presented here show:

1. Reads originating from targeted genes are depleted at > 90%.

2. CRISPRclean depletion results in higher numbers of novel in catalogue and novel not in catalogue isoforms.

3. Total isoforms generated increased by 375%.

4. Data from the depleted sample contained 7.23-fold more unique molecules, 1.9-fold more unique genes, and 4.5-fold more unique isoforms than the control.

To learn more, visit jumpcodegenomics.com

Ordering information

Catalog	Product name	Samples
KIT1018	CRISPRclean Single Cell RNA Boost Kit	24

References

1. Petti AA, Williams SR, Miller CA, Fiddes IT, Srivatsan SN, Chen DY, Fronick CC, Fulton RS, Church DM, Ley TJ. A general approach for detecting expressed mutations in AML cells using single cell RNA-sequencing. Nat Commun. 2019 Aug 14;10(1):3660.

2. Moreno-Villanueva M, Zhang Y, Feiveson A, Mistretta B, Pan Y, Chatterjee S, Wu W, Clanton R, Nelman-Gonzalez M, Krieger S, Gunaratne P, Crucian B, Wu H. Single-Cell RNA-Sequencing Identifies Activation of TP53 and STAT1 Pathways in Human T Lymphocyte Subpopulations in Response to Ex Vivo Radiation Exposure. Int J Mol Sci. 2019 May 10;20(9):2316.

3. Huang KK, Huang J, Wu JKL, Lee M, Tay ST, Kumar V, Ramnarayanan K, Padmanabhan N, Xu C, Tan ALK, Chan C, Kappei D, Göke J, Tan P. Long-read transcriptome sequencing reveals abundant promoter diversity in distinct molecular subtypes of gastric cancer. Genome Biol. 2021 Jan 22;22(1):44.

4. Veiga DFT, Nesta A, Zhao Y, Mays AD, Huynh R, Rossi R, Wu TC, Palucka K, Anczukow O, Beck CR, Banchereau J. A comprehensive long-read isoform analysis platform and sequencing resource for breast cancer. Sci Adv. 2022 Jan 21;8(3).

5. Palmer CR, Liu CS, Romanow WJ, Lee MH, Chun J. Altered cell and RNA isoform diversity in aging Down syndrome brains. Proc Natl Acad Sci U S A. 2021 Nov 23;118(47).

6. Tardaguila M, de la Fuente L, Marti C, Pereira C, Pardo-Palacios FJ, Del Risco H, Ferrell M, Mellado M, Macchietto M, Verheggen K, Edelmann M, Ezkurdia I, Vazquez J, Tress M, Mortazavi A, Martens L, Rodriguez-Navarro S, Moreno-Manzano V, Conesa A. SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. Genome Res. 2018 Feb 9;28(3):396–411. doi: 10.1101/gr.222976.117. Epub ahead of print. Erratum in: Genome Res. 2018 Jul;28(7):1096. PMID: 29440222; PMCID: PMC5848618.