Highly efficient full-length isoform sequencing using intramolecular multiplexed cDNAs: a 16-22x yield increase on PacBio sequencers

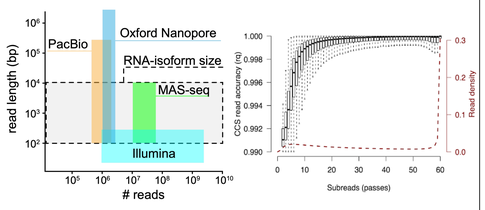
Authors: Aziz Al'Khafaji*[1], **Jonathan T. Smith***[1], Kiran V Garimella*[1], Mehrtash Babadi*[1], Maura Costello[1], Michael Gatzen[1], Ally Day[1], Stacey Gabriel[1], Eric Banks[1], Anthony Philippakis[1], Paul Blainey[1,2], Nir Hacohen[1,2,3,4]
Affiliations: 1 - The Broad Institute of MIT and Harvard, Cambridge MA 02142; 2 - Massachusetts Institute of Technology, Cambridge, MA, 02142; 3 - Harvard Medical School, Boston, MA 02115; 4 - Massachusetts General Hospital, Charlestown, MA 02129; * - *Authors contributed equally*

Abstract:

High-throughput full-length RNA isoform sequencing is currently cost-prohibitive, constraining our ability to understand the transcriptional diversity that drives and regulates dynamic and heterogeneous biological systems. Here, we introduce, validate and apply a novel intramolecular cDNA multiplexing approach, *MAS-seq*, that boosts full-length RNA isoform sequencing output >15 fold to approximately 40 million cDNA reads per run on a long-read sequencing platform. We demonstrate that this added sequencing throughput drives robust cell clustering and vastly enhances both isoform quantification and discovery of differentially spliced genes.

Introduction

Scalable full-length RNA isoform identification and quantification remain elusive goals for bulk and single-cell studies as the necessary read lengths (>5 kb) and depths (>2x10^7 reads) are not easily attainable by existing sequencing platforms. For example, short-read sequencing platforms (e.g. Illumina) achieve more than sufficient throughput (>1x10^9 reads) but are hindered by limited read lengths (50 - 600 bp) which are inadequate to span the vast majority of human transcripts (~ 1.6 +/- 1.1 kb).

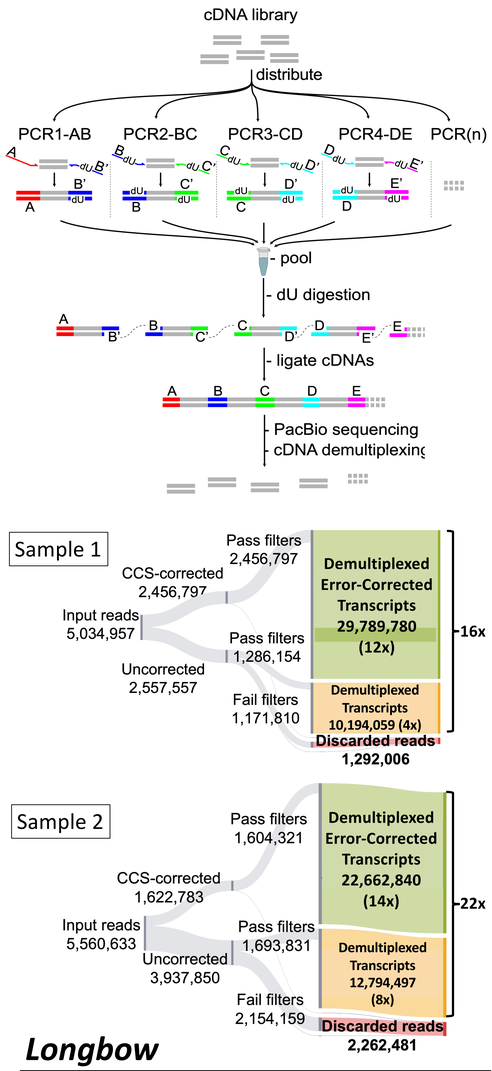


On the PacBio Sequel II platform, consensus base quality reaches Phred-scale quality of ~Q30 around 10 circular passes, with subsequent consensus reads providing only nominal utility. For the current Sequel II instrument and SMRT Cell 8M chemistry, 15 - 20 kb is the optimal library size for reaching ~10 circular passes. As the length register of transcriptomic sequences is on average substantially shorter (100 bp - 5 kb), the number of circular passes is consequentially much higher (50 - 60), wasting sequencing capacity

MAS-seq

To maximize the sequencing potential on the PacBio platform, we have developed an unbiased method for programmatic cDNA concatemerization, **M**ultiplexed **A**rray**s** sequencing (***MAS-seq***). Through the use of deoxy-uracil digestion followed by deterministic barcode-directed ligation of cDNAs, MAS-seq generates long multiplexed cDNA arrays with a narrow length distribution that allows for both accurate consensus sequencing and more optimal capacity utilization.

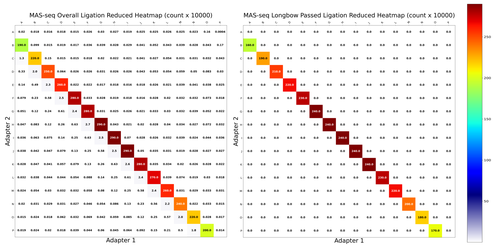
In combination with upstream artifact depletion measures, MAS-seq boosts the sequencing throughput to approximately 40 million full-length transcripts per SMRT Cell 8M flow cell, a >15-fold increase over CCS corrected read counts.



Longbow

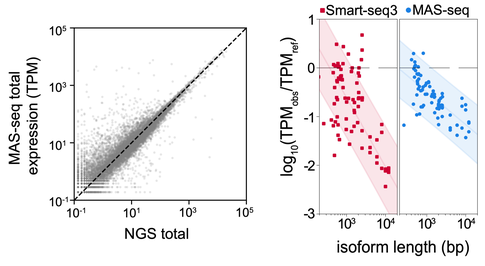
The fixed pattern of distinct MAS-seq adapters provides landmarks for effective cDNA segmentation as well as constraints for detecting malformed or otherwise defective sequences. To exploit these signals, we developed a composite profile hidden Markov model, ***Longbow***, for the probabilistic annotation and optimal segmentation (via maximum *a posteriori* state path) of each MAS-seq read. In this formulation, a MAS-seq read is considered to be a mosaic of imperfect (but complete) copies of the various known adapter sequences among which the unknown cDNA sequences of interest are present.

Longbow is robust to the presence of a high per-base error rate. On average across our MAS-seq libraries, ~74.3% of reads (99.2% CCS corrected, 52.3% CCS uncorrected) were found to segment correctly. Segmentation results inconsistent with our expected array structure (i.e. off-subdiagonal elements of the matrices below) were filtered out (33.17% of arrays).

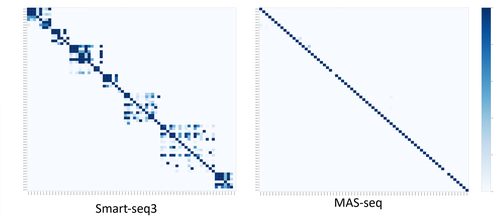


Validation

We performed full-length RNA sequencing on the Spike-In RNA Variants (SIRV-set 4), containing 69 synthetic RNA isoforms of varying lengths and equal molarity, and 92 ERCC RNA standards whose concentration span 6 orders of magnitude (Lexogen). Smart-seq3 sequencing of the SIRV-set 4 standards was performed in parallel to compare short-read isoform reconstructions to the high-throughput long-read sequencing approach. Quantification results were broadly similar overall between both protocols.



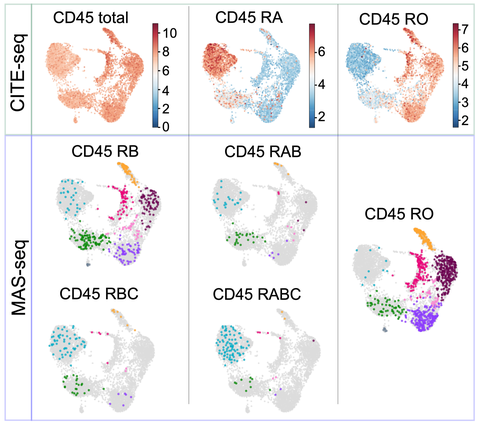
Characterization of long isoforms was markedly improved in MAS-seq and Iso-Seq versus Smart-seq3. Smart-seq3 isoform reconstructions also exhibited substantial ambiguity (~43%) in assigning reconstructions to a specific known isoform, often exhibiting confusion among isoforms derived from the same synthetic SIRV gene. In contrast, MAS-seq data provide direct identification of transcript isoforms without the need for reconstruction, and hence assign isoform identities with nominal ambiguity (~0.004).



CD45 Isoform Characterization

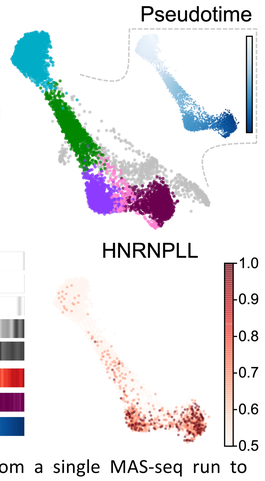
To characterize MAS-seq performance on single-cell RNA sequencing libraries, we performed 10x Genomics 5' single-cell gene expression on tumor-infiltrating CD8+ T cells. From the same 10x full-length cDNA library we generated both standard short-read and MAS-seq long-read libraries. Despite large discrepancies in sequencing depth between short and long-read approaches, cellular clustering was highly similar.

A common set of T cell transcriptional states ranging from stem cell-like to terminally differentiated were observed. Leveraging the canonical and distinct splicing patterns of CD45 over the course of T cell differentiation, we performed orthogonal validations of CD45 isoform expression at the protein level using CITE-seq and the mRNA level using MAS-seq. CD45 isoform expression between these two modalities was highly concordant.

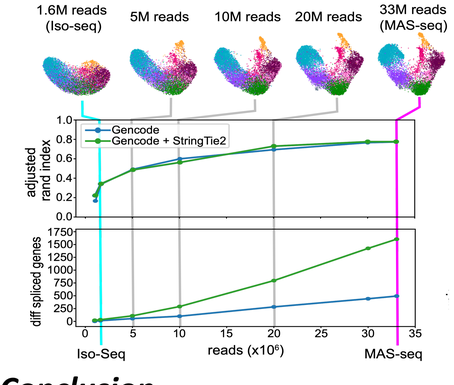


Notably, mRNA measurements were more granular in their ability to resolve the many CD45 isoforms present (RO, RA, RAB, RB, RBC) as compared to the antibody-based CITE-seq approach. This is due to the single epitope specificity antibodies which do not enable discrimination of closely related isoforms.

Canonical CD45 isoform expression and its associated splicing factor, HNRNPLL, tracked clearly along this differentiation trajectory.



We subsampled reads from a single MAS-seq run to relevant counts and computed the adjusted rand index (ARI) and number of differentially spliced genes. We observed a 44% gain in ARI in single-cell clustering and a 34-fold gain in identifying differentially spliced genes between CD8+ T cell subtypes (multiple hypothesis testing correction with FDR < 0.05)



Conclusion

With its compatible nature, MAS-seq and Longbow are poised to facilitate isoform discovery and reference generation with cell type annotations at scale.