Single Cell Isoform Sequencing (sclso-Seq) Identifies Novel Full-length mRNAs and Cell Type-specific Expression

Elizabeth Tseng[1], **Jason G. Underwood**[1], Aparna Bhuduri[2], Calum Marrs[3], Filip Konopacki[3], Daniel Wong[3], Katherine M. Munson[5], Alexandra Lewis[5], Alex A. Pollen[2] and Evan E. Eichler[4,5]

[1] PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025 [2] Department of Neurology, University of California-San Francisco, San Francisco, CA
[3] Dolomite/Blacktrace Holdings Ltd; 27 Jarman Way, Royston, SG8 5HW, UK [4] Howard Hughes Medical Institute, University of Washington, Seattle, WA
[5] Department of Genome Sciences, University of Washington, Seattle, WA

Abstract

Single cell RNA-seq (scRNA-seq) is an emerging field for characterizing cell heterogeneity in complex tissues. However, most scRNA-seq methodologies are limited to gene count information due to short read lengths.

Here, we combine the microfluidics scRNA-seq technique, Drop-Seq, with PacBio Single Molecule, Real-Time (SMRT) Sequencing to generate full-length transcript isoforms that can be confidently assigned to individual cells.

We generated single cell Iso-Seq (scIso-Seq) libraries for chimp and human cerebral organoid samples on the Dolomite Nadia platform and sequenced each library with two SMRT Cells 8M on the PacBio Sequel II System. We developed a bioinformatics pipeline to identify, classify, and filter full-length isoforms at the single-cell level. We show that scIso-Seq reveals full-length isoform information not accessible using short reads that can reveal differences between cell types and amongst different species.

Single Cell Iso-Seq (sclso-Seq)

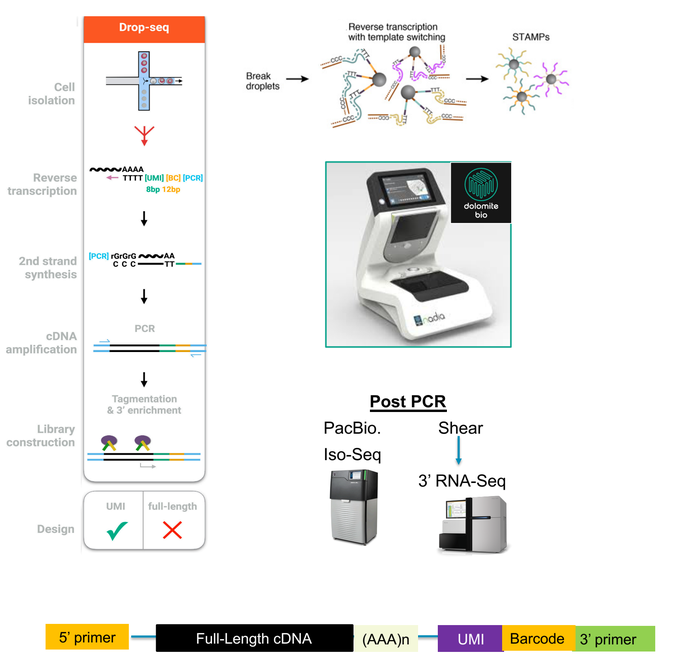


Figure 1. Full-length single cell libraries were generated using the Dolomite Nadia system and made into PacBio Iso-Seq libraries. Sequencing was performed on the Sequel II System.

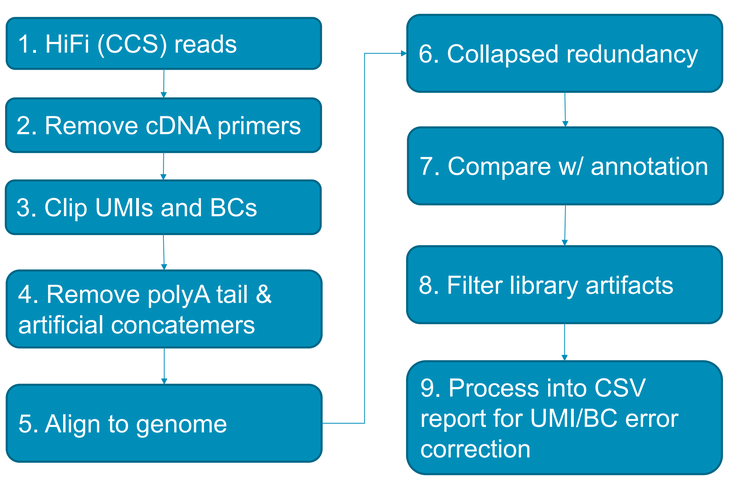


Figure 2. Bioinformatics analysis for scIso-Seq data. The pipeline is described in Cupcake [1].

sclso-Seq on Cerebral Organoids

SAMPLE	POL READS (bp)	POL BASE (GB)	POL LENGTH (kb)
Chimp Organoid	3,157,575	199	62.9
Chimp Organoid	3,637,178	206	56.7
Human Organoid	2,856,715	177	62.1
Human Organoid	5,277,118	265	50.2

Table 1. Sequencing yield for the chimp and human organoid single cell Iso-Seq (scIso-Seq) libraries on each of two SMRT Cells 8M run on the Sequel II System.

SAMPLE	FLNC (filtered)	UNIQUE READS	UNIQUE GENES	UNIQUE ISOFORMS
Chimp Organoid	2,303,267	418,542	14,049	58,892
Human Organoid	2,291,947	382,734	14,737	60,815

Table 2. Number of unique (de-duplicated) full-length, non-concatemer (FLNC) reads and corresponding number of unique genes and transcript isoforms.

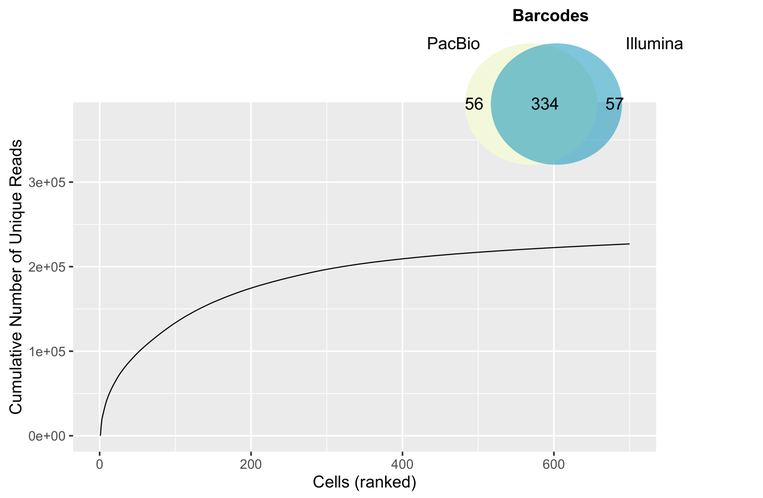


Figure 3. Matching cell barcodes between PacBio scIso-Seq data and Illumina short-read data shows high concordance. The scIso-Seq cumulative read plot indicates ~400 STAMPS (single cells).

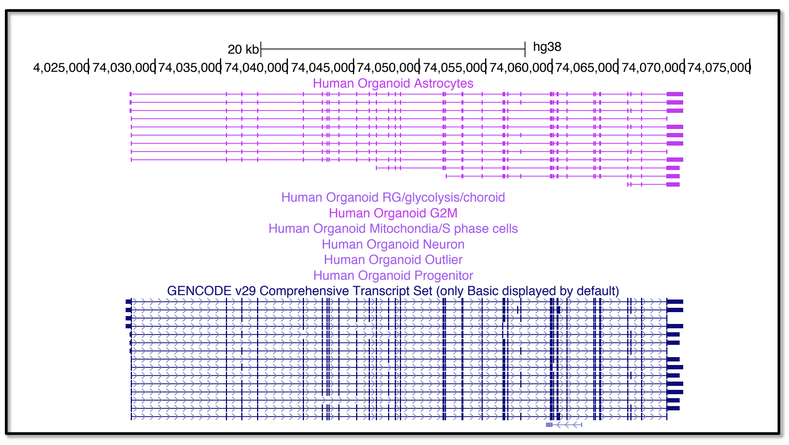


Figure 4. The tropoelastin gene is found only to be expressed in astrocytes but not other human organoid cell types and shows exon skipping events and usage of alternative start/end sites.

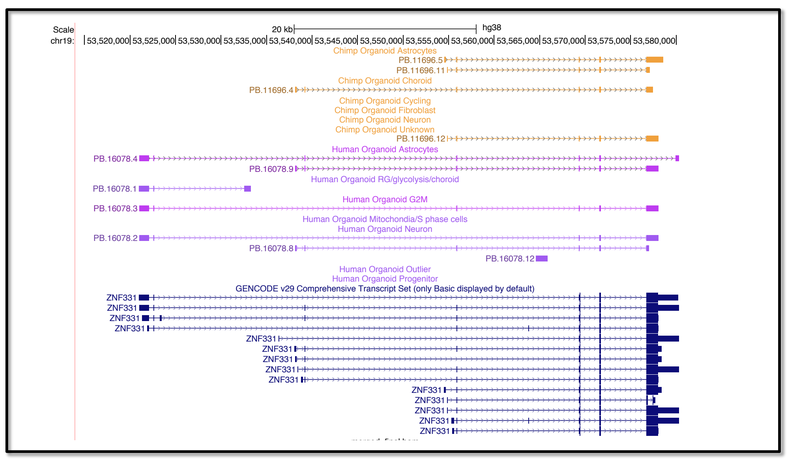
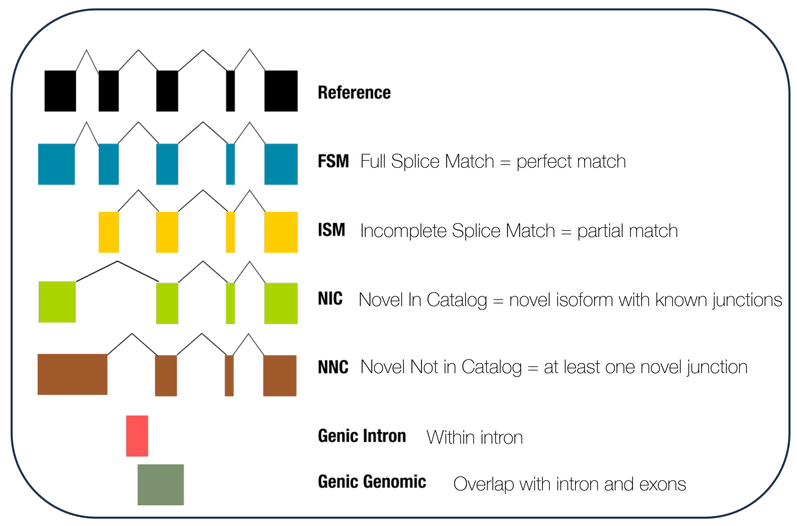


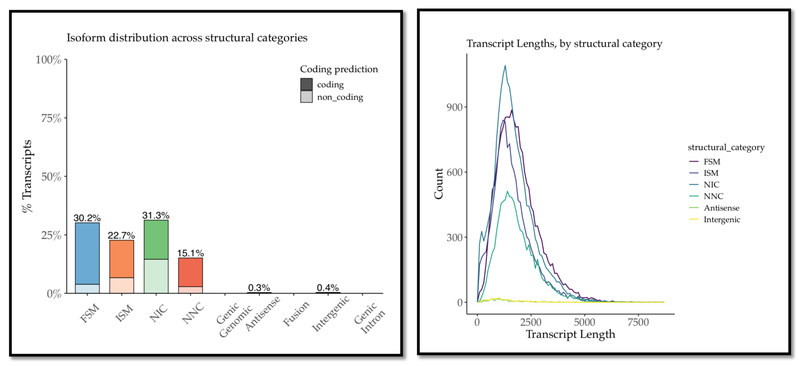
Figure 5. Alternative transcription start site (TSS) usage in chimp and human in ZNF331.

Using SQANTI2 for QC

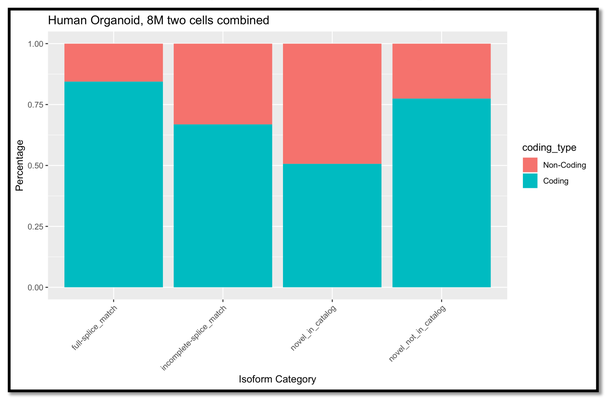
SQANTI2 [2] matching Iso-Seq transcripts to existing annotations and provides a wide range of descriptors of transcript quality.



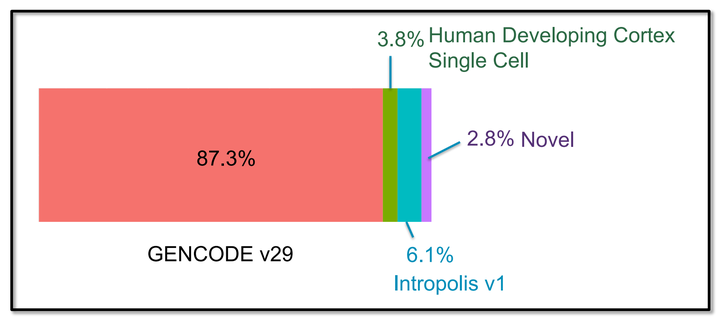
(a) Comparison against Gencode v29 transcript annotation reveals ~46% scIso-Seq transcripts are novel.



(b) Novel scIso-Seq isoforms retain coding potential



(c) Most junctions are validated by existing annotations and RNA-seq data.



(d) Matching FANTOM5 CAGE peak data with scIso-Seq transcripts shows full splice matches (perfect junction matches to annotations) are enriched for known TSS.

Class	# Isoforms	% with CAGE Peak ≤50 bp
Full Splice Match	18,344	78%
Incomplete Splice Match	13,802	37%
Novel In Catalog	19,033	44%
Novel Not In Catalog	9,197	67%
Intergenic	245	29%

Figure 6. SQANTI2 comparison against existing annotations confirms scIso-Seq data consists of full-length (5'-3') transcript isoforms.

REFERENCES

[1] cDNA_Cupcake https://github.com/Magdoll/cDNA_Cupcake
[2] SQANTI2 https://github.com/Magdoll/SQANTI2/