From RNA to Full-Length Transcripts: The PacBio Iso-Seq Method for Transcriptome Analysis and Genome Annotation

Michelle N. Vierra[1], Sarah B. Kingan[1], Elizabeth Tseng[1], Ting Hon[1], William J. Rowell[1], Jacquelyn Mountcastle[2], Olivier Fedrigo[2], Erich D. Jarvis[2], Jonas Korlach[1]

1. PacBio, Menlo Park, CA 94025. 2. The Rockefeller University, New York, NY & Howard Hughes Medical Institute.

Abstract

A single gene may encode a surprising number of proteins, each with a distinct biological function. This is especially true in complex eukaryotes. Short-read RNA sequencing (RNA-seq) works by physically shearing transcript isoforms into smaller pieces and bioinformatically reassembling them, leaving opportunity for misassembly or incomplete capture of the full diversity of isoforms from genes of interest.

The PacBio Isoform Sequencing (Iso-Seq™) method employs long reads to sequence transcript isoforms from the 5' end to their poly-A tails, eliminating the need for transcript reconstruction and inference. These long reads result in complete, unambiguous information about alternatively spliced exons, transcriptional start sites, and poly-adenylation sites. This allows for the characterization of the full complement of isoforms within targeted genes, or across an entire transcriptome.

Here we present improved genome annotations for two avian models of vocal learning, Anna's hummingbird (*Calypte anna*) and zebra finch (*Taeniopygia guttata*), using the Iso-Seq method. We present graphical user interface and command line analysis workflows for the data sets. From brain total RNA, we characterize more than 15,000 isoforms in each species, 9% and 5% of which were previously unannotated in hummingbird and zebra finch, respectively. We highlight one example where capturing full-length transcripts identifies additional exons and UTRs.

Sample Preparation & Sequencing

TISSUE DISSECTION
- Birds dissected immediately post-mortem
- Tissue stored in a cryogenic tube or embedded in OCT resin at -80°C

TOTAL RNA ISOLATION
- 30 mg whole brain used with Qiagen RNeasy Mini Kit
- 6.76 µg (zebra finch) & 3.38 µg (hummingbird) of total RNA extracted

BARCODED 1ST STRAND cDNA
- 1 µg of each RNA sample used in RT rxns
- Barcoded oligo-dT used for first strand cDNA synthesis

PCR OPTIMIZATION
- A single PCR with aliquots removed at 8, 10, 12, and 14 cycles was performed for each sample
- 11 cycles determined as optimal

AMPLIFIED CDNA
- cDNA amplified for 11 cycles according to PCR procedures in the Iso-Seq Template Preparation protocol[1] and optimization

Size Selection (optional)
- A portion of the amplified cDNA from each sample was enriched for 5-10 kb transcripts and pooled with non size-selected cDNA

SMRTbell™ LIBRARY PREPARATION
- PCR products purified
- DNA damage repair and end repair was performed
- SMRTbell adaptors ligated to DNA

- Primer annealed and polymerase bound to templates according to the binding calculator
- Barcoded samples run on 4 Sequel 1M SMRT cells with 2.0 chemistry

Figure 1. Sample Prep & Sequencing Workflow for Iso-Seq Method on the Sequel System.

Sequencing Results

SMRT Cell ID	Reads	Yield	Polymerase Read N50	Subread N50
1	507,715	7.7 Gb	35 kb	5.8 kb
2	517,278	7.4 Gb	32 kb	5.8 kb
3	464,945	7.2 Gb	35 kb	5.3 kb
4	347,004	6.1 Gb	38 kb	5.3 kb

Table 1. Performance Statistics for Sequencing. Barcoded libraries were constructed for zebra finch and hummingbird from brain total RNA and run with Sequel 2.0 chemistry.

Iso-Seq Analysis in SMRT Link 5.0

DATASET IMPORT
- Dataset of subreads; can contain multiple SMRT cell runs

CCS
- Generation of circular consensus sequences

CLASSIFY
- Classify CCS reads into full length (FL) and non-full length (nFL) reads based on presence of cDNA primers and polyA tail

CLUSTER
- Isoform-level clustering of FL reads to generate initial consensus

POLISH
- Include nFL reads in clusters results and final consensus calling with Arrow

MAP
- Map isoforms sequences to genome using GMAP and remove redundancy

FINAL OUTPUT
- Annotated isoforms in GFF
- Isoforms sequences in FASTQ
- Read support for each isoform
- Additional Files

Figure 2. SMRT Link 5.0 Iso-Seq with Mapping Workflow. Isoforms were mapped to primary contigs of PacBio FALCON-Unzip assemblies for zebra finch (GCA_002008985.2) and hummingbird (GCA_002021895.1).

Command Line Analysis with Iso-Seq2

Data were analyzed with Iso-Seq2, an advanced version of Iso-Seq that is currently available in beta release[2,3] and will be incorporated into future SMRT Link releases.

IsoSeq Classify

CLASSIFY
- INPUT: ccs.bam or subreads.bam
- OUTPUT: isoseq_flnc.fa, isoseq_nfl.fa

ToFU2

PRECLUSTER with minimap
- INPUT: isoseq_flnc.fa
- OUTPUT: preCluster_out/<bin>

ICE2 iterative clustering
- *Within each <bin> (n=10000):*
- INPUT: isoseq_flnc.fa
- OUTPUT: final_consensus.fa

COLLECT AND CHUNK
- INPUT :preCluster_out/<bin>
- OUTPUT: collected_final.chunk1.fa,..., collected_final.chunk20.fa

ICE_ARROW2
- INPUT:collected_final.chunk*.fa, isoseq_nfl.fa
- OUTPUT:[hq|lq].polished.fq

Cupcake

GMAP
- INPUT: genome_reference.fa, hq_isoforms.polished.fq
- OUTPUT:hq_isoforms.fq.sam

COLLAPSE and FILTER
- INPUT:hq_isoforms.fq.sam
- OUTPUT:final_hq_isoforms.fq

Figure 3. Command line Iso-Seq analysis workflow. Summary of analysis modules, file inputs and outputs[2,3].

Isoform Analysis Results

	Zebra Finch	Hummingbird
HQ isoforms	17,451	16,944
Mean length	4,112 bp	3,938 bp
Number of loci	7,258	7,418
Mean isoforms per locus	2.40	2.28
Novel isoforms	901 (5.2%)	1,517 (9.0%)
Mapped back to PacBio reference	17,450 (99%)	16,944 (100%)
Mapped to Previous Reference[4,5]	16,183 (93%)	14,608 (86%)

Table 2. Isoform Characterization and Mapping with ToFU2 and Cupcake. High Quality (HQ): ≥ 2 full length reads, >99% accuracy. Isoforms belong to same locus if they overlap by at least 1 bp on same strand. Novel isoforms lack BLAST hit to current transcriptome (GCF_000151805.1, GCF_000699085.1) with e-value < 1^{-10}. HQ Isoforms mapped to references with GMAP filters: min. coverage 90%, min. identity 95%.

Example of Full-Length Transcript Capture: New Exons and UTRs

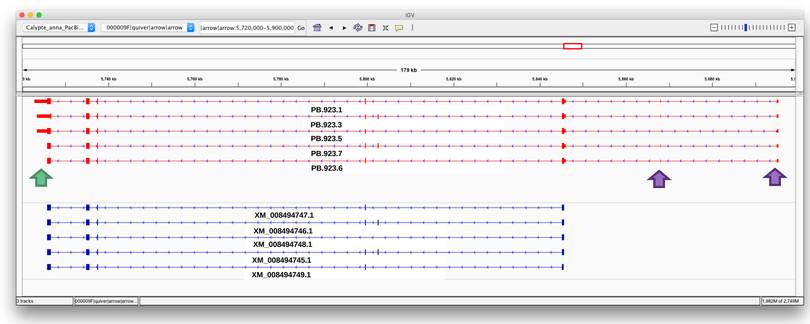

Figure 4. Multiple Isoforms for neuroligin-4 in hummingbird. Full- length isoform sequences (red transcript models) identify two additional non-coding 5' exons (purple arrows) and extended 3' UTRs (green arrow), while also capturing all five known splice variants (blue transcript models). Neuroligins are a class of postsynaptic cell-adhesion molecules essential for normal synapse function; neuroligin-4 is implicated in autism spectrum disorder[6]. Isoforms visualized in IGV v3.0_beta.

Conclusion

PacBio Iso-Seq method for sequencing RNA is effective for the survey of transcript diversity to improve genome annotation and gene discovery. We demonstrated:

- Improved loading on Sequel System simplifies library prep and makes size selection optional
- Both SMRT Link and command line tools can be used to perform analyses
- Significant numbers of novel isoforms can be characterized, even in model organisms
- Full-length transcripts provide new exon and UTR sequences

References

1. http://www.pacb.com/wp-content/uploads/Procedure-Checklist-Iso-Seq-Template-Preparation-Sequel-Systems.pdf
2. Tseng E. (2017) https://github.com/PacificBiosciences/IsoSeq_SA3nUP/
3. Tseng E. (2017) https://github.com/Magdoll/cDNA_Cupcake/
4. Warren WC, et al. (2010). The genome of a songbird. *Nature* 464(7289), 757-762.
5. Zhang G, et al. (2014). Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* 346(6215),1311-1320.
6. Südhof TC (2008) Neuroligins and neurexins link synaptic function to cognitive disease. *Nature* 455(7215), 903–911.

Acknowledgements

The authors would like to thank Yuan Li and Emily Hatas for their assistance with this project.