A Complete Solution for High-Quality Genome Annotation Using the PacBio Iso-Seq Method

Elizabeth Tseng

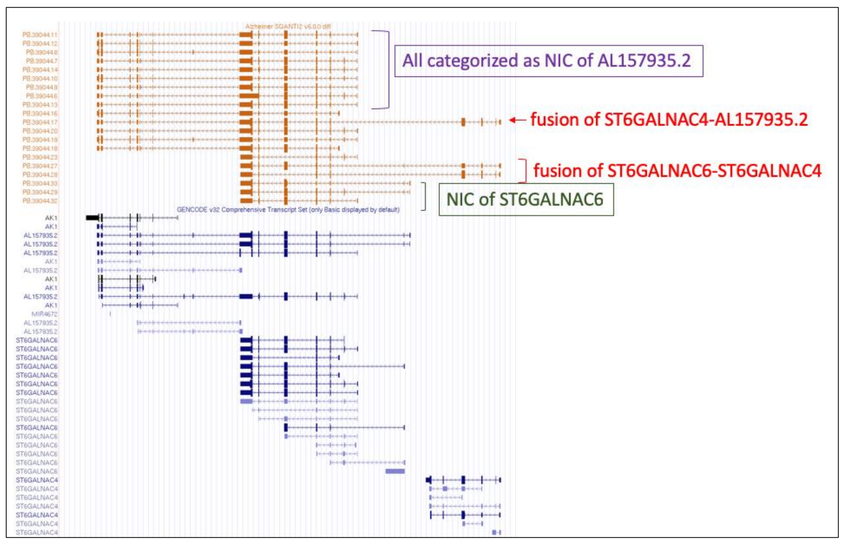
PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025

RNA SEQUENCING

Iso-Seq on the Sequel II System

- Generate full-length transcripts of 10 kb or longer
- High accuracy (>99%) for ORF prediction
- No reference genome required
- Bioinformatics tools from raw data to functional annotation
- Many applications, including: [1]
  - Genome annotation
  - Novel gene and isoform discovery
  - Fusion gene detection
  - Allele-specific isoform expression analysis
  - Better reference for RNA-seq quantification
  - Assess genome assembly quality
  - Single cell analysis

SQANTI2 for Complex Transcript Classification

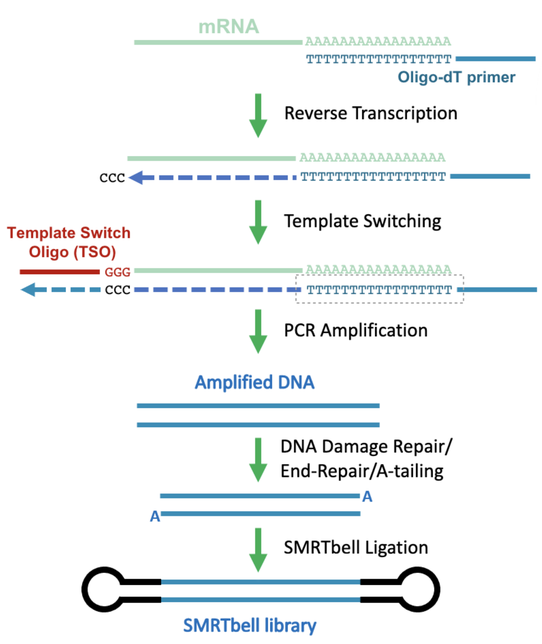


SQANTI2 classifies isoforms from an Alzheimer brain Iso-Seq dataset at a complex locus.

Novel isoforms (NIC) and readthrough transcripts of multiple genes identified.

Iso-Seq Express Kit [2]

- Input 60-300 ng total RNA
- Full-length cDNA
- Multiplexing support



LIBRARY PREP
1 DAY

SMRT SEQUENCING
1 DAY

DATA ANALYSIS
1 DAY

Sequel II System

- 1 SMRT Cell 8M for whole transcriptome
- Up to 4 million full-length reads

	FL Reads	Unique Genes	Unique Transcripts
UHRR [3]	4,734,362	16,328	183,689
Alzheimer Brain [4]	4,277,293	17,670	162,290

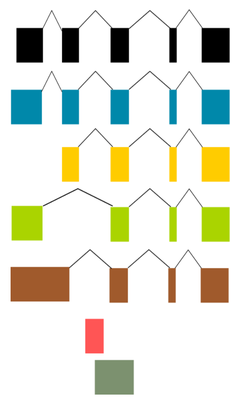
Main Bioinformatics Tools

	Input	Output
Sequencing		`subreads.bam`
Iso-Seq Analysis	`subreads.bam` or `ccs.bam`	`Collapsed unique transcripts (GFF, FASTA)`

Transcript Classification	`Unique transcripts`
	`Reference genome`
	`Annnotation (GTF)`
	`CAGE Peak`
	`Junction data...`
		`Transcript classification`
		`Junction classification`
		`Figures`

Functional Annotation	`SQANTI output`	`Annotated GTF`
Differential Analysis	`Experimental design` `Annotated GTFs`	Isoform Expression / Time ----->

SQANTI2 Transcript Classification



Full Splice Match, matches reference perfectly.

Incomplete Splice Matches, matches reference partially

Novel In Catalog, novel isoform using known junctions

Novel Not in Catalog, novel isoforms using novel junctions.

Supporting Bioinformatics Tools

Cupcake
- collapse redundant transcripts
- merge multi-sample output
- saturation curve
- file format conversion
- single cell analysis

CONGENT (Coding Genome reconstruction Tool)
- gene family finding
- genome reconstruction
- evaluate assembly

TAMA GO
- collapse redundant transcripts
- merge multi-sample output
- NMD/ORF prediction
- transcript filtering

TALON
- long read processing & annotation pipeline developed independently by ENCODE4

[1] Low et al., "**Haplotype-Resolved Cattle Genomes Provide Insights Into Structural Variation and Adaptation**", *biorxiv* (2019)
Beiki et al., "**Improved annotation of the domestic pig genome through integration of Iso-Seq and RNA-seq data**", *BMC Genomics* (2019)
Wang et al., "**Reviving the Transcriptome Studies: An Insight Into the Emergence of Single-Molecule Transcriptome Sequencing**", *Front Genet* (2019)
[2] Iso-Seq Express: https://www.pacb.com/wp-content/uploads/Procedure-Checklist-Iso-Seq-Express-Template-Preparation-for-Sequel-and-Sequel-II-Systems.pdf
[3] UHRR: https://github.com/PacificBiosciences/DevNet/wiki/Sequel-II-System-Data-Release:-Universal-Human-Reference-(UHR)-Iso-Seq
[4] Alzheimer brain: https://downloads.pacbcloud.com/public/dataset/Alzheimer2019_IsoSeq/