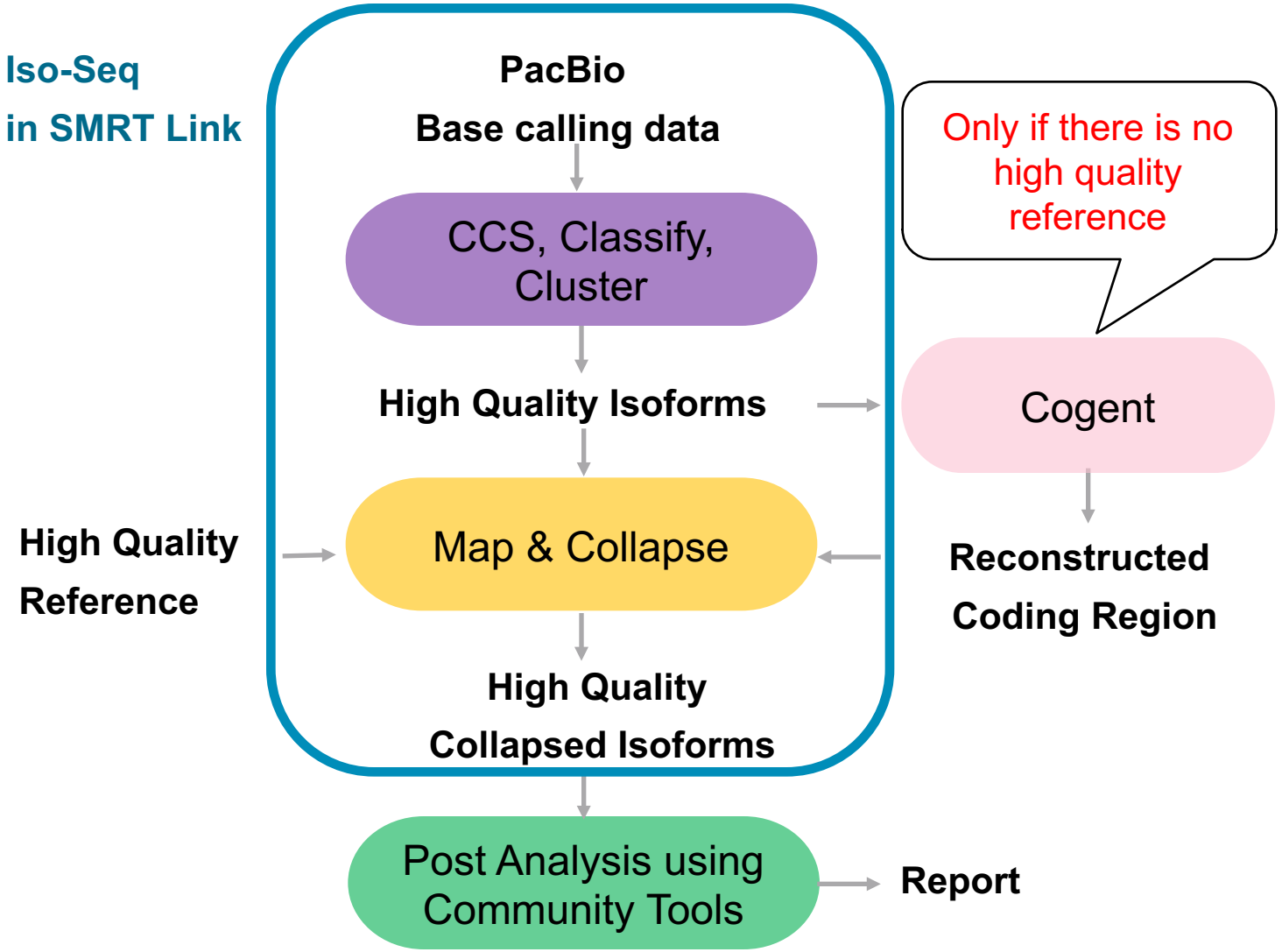
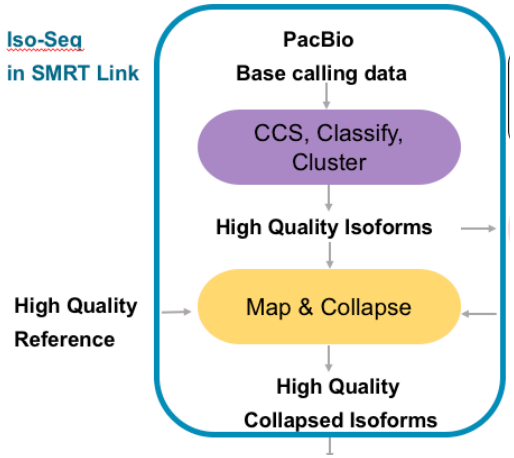
Iso-Seq Deep Dive

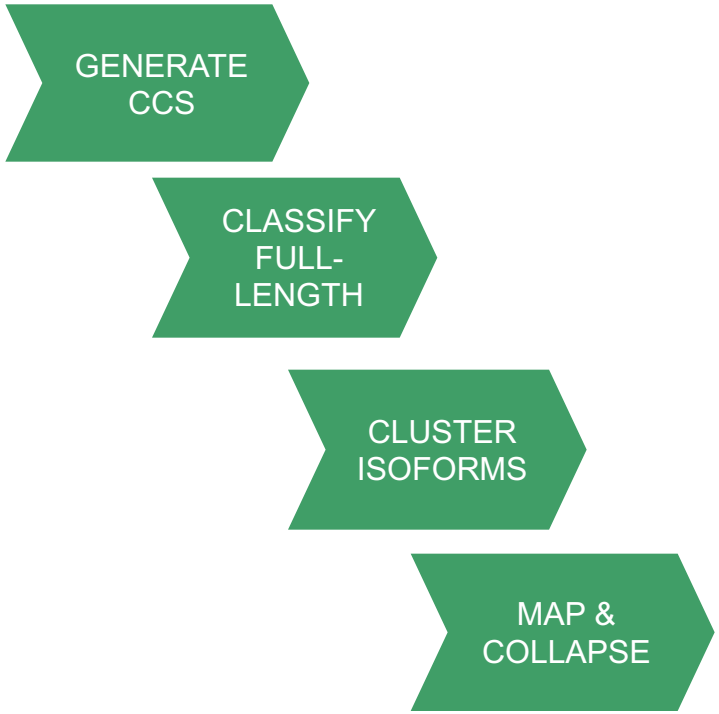
Elizabeth Tseng & Yuan Li & Armin Töpfer, May 2018

ISO-SEQ ANALYSIS WORKFLOW

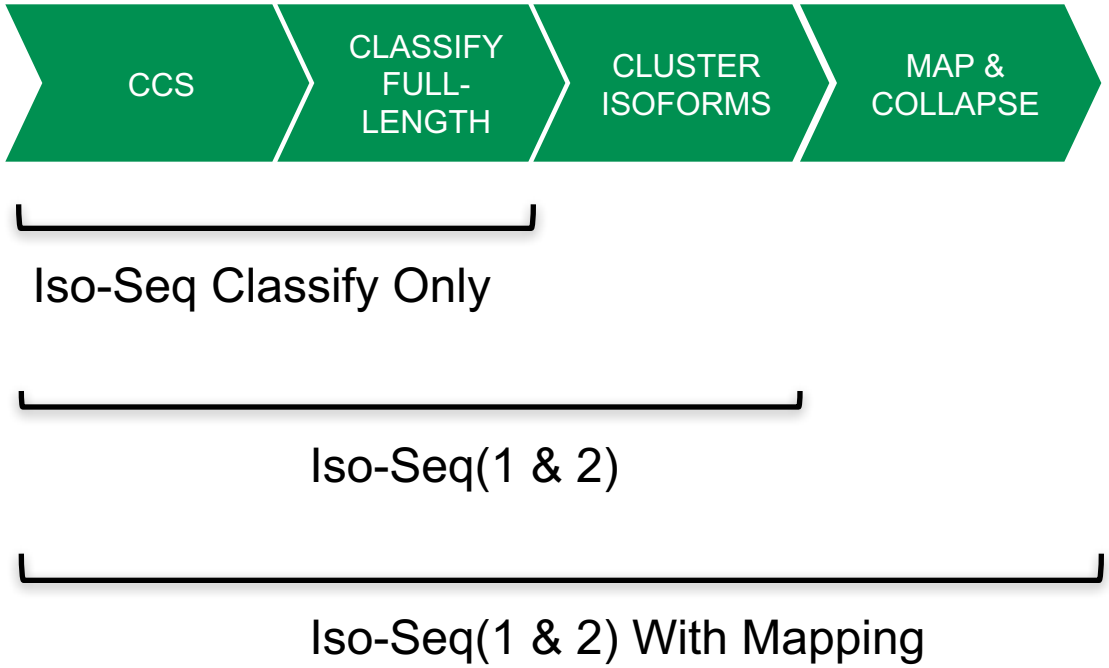


ISO-SEQ: FULL-LENGTH TRANSCRIPT SEQUENCING

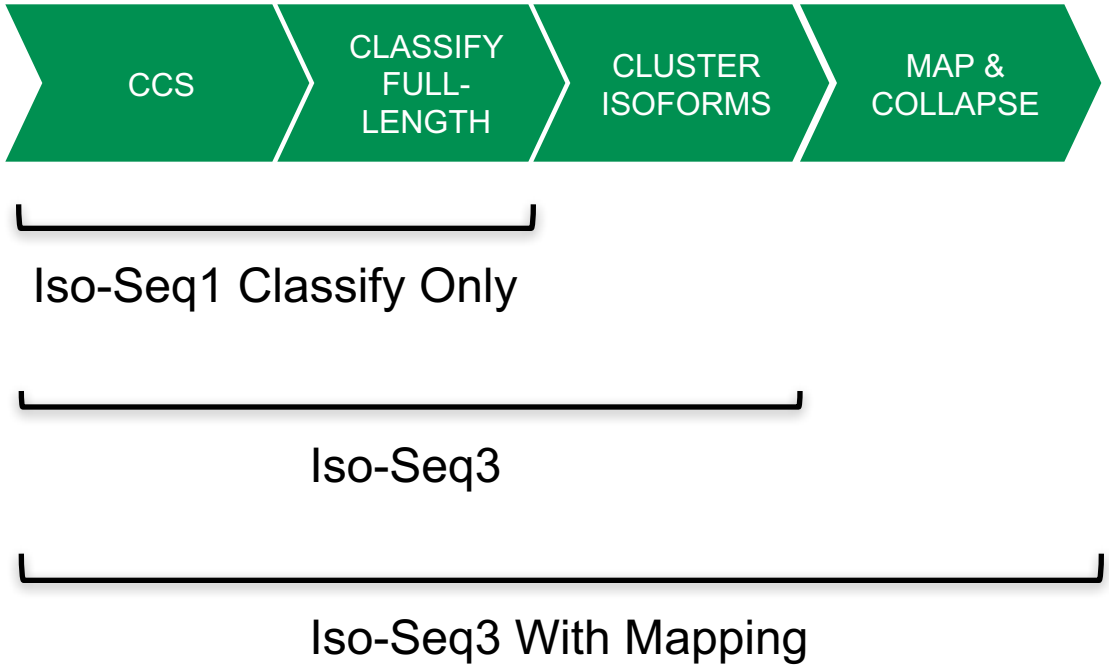




CURRENT: SMRT LINK 5.1



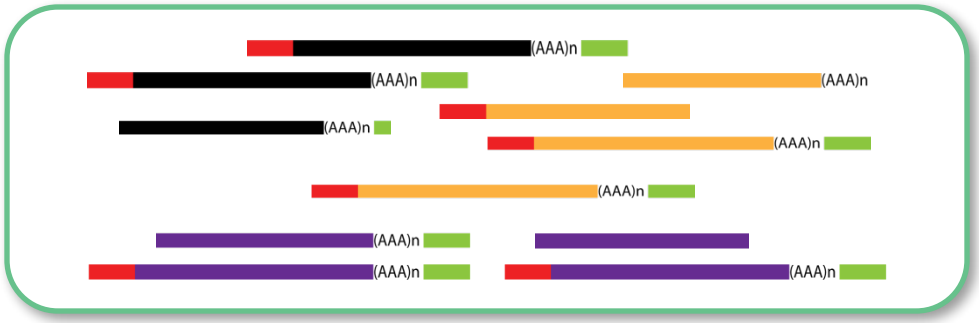
FUTURE: SMRT LINK V6.0

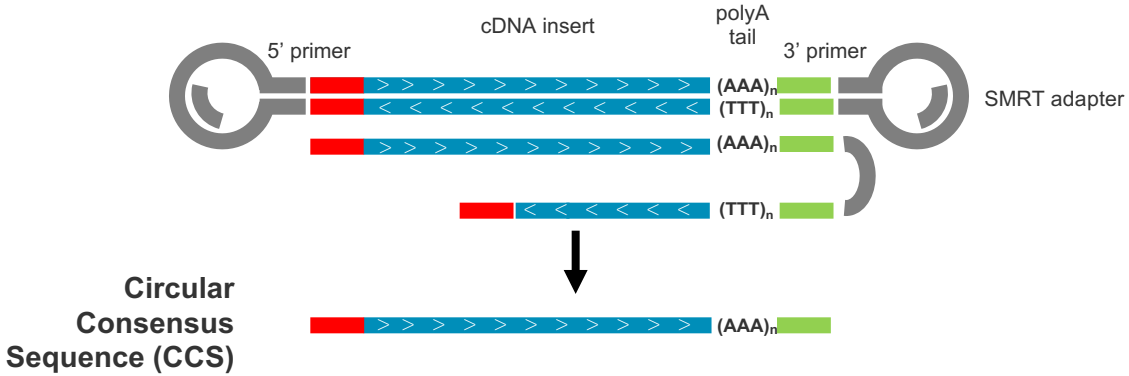


- Iso-Seq1 and Iso-Seq1 With Mapping will be obsolete in the future
- Iso-Seq2 and Iso-Seq2 with Mapping will be removed in SMRT Link v6.0

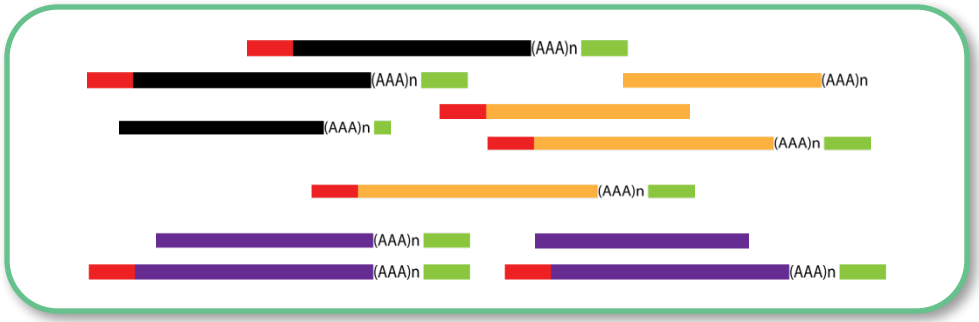

CCS
CLASSIFY FULL-LENGTH
CLUSTER ISOFORMS CLUSTER
MAP & COLLAPSE

CCS


(AAA)n


5' primer
cDNA insert
polyA tail
3' primer
SMRT adapter
(AAA)n
(TTT)n
Circular Consensus Sequence (CCS)

CCS



nFL reads

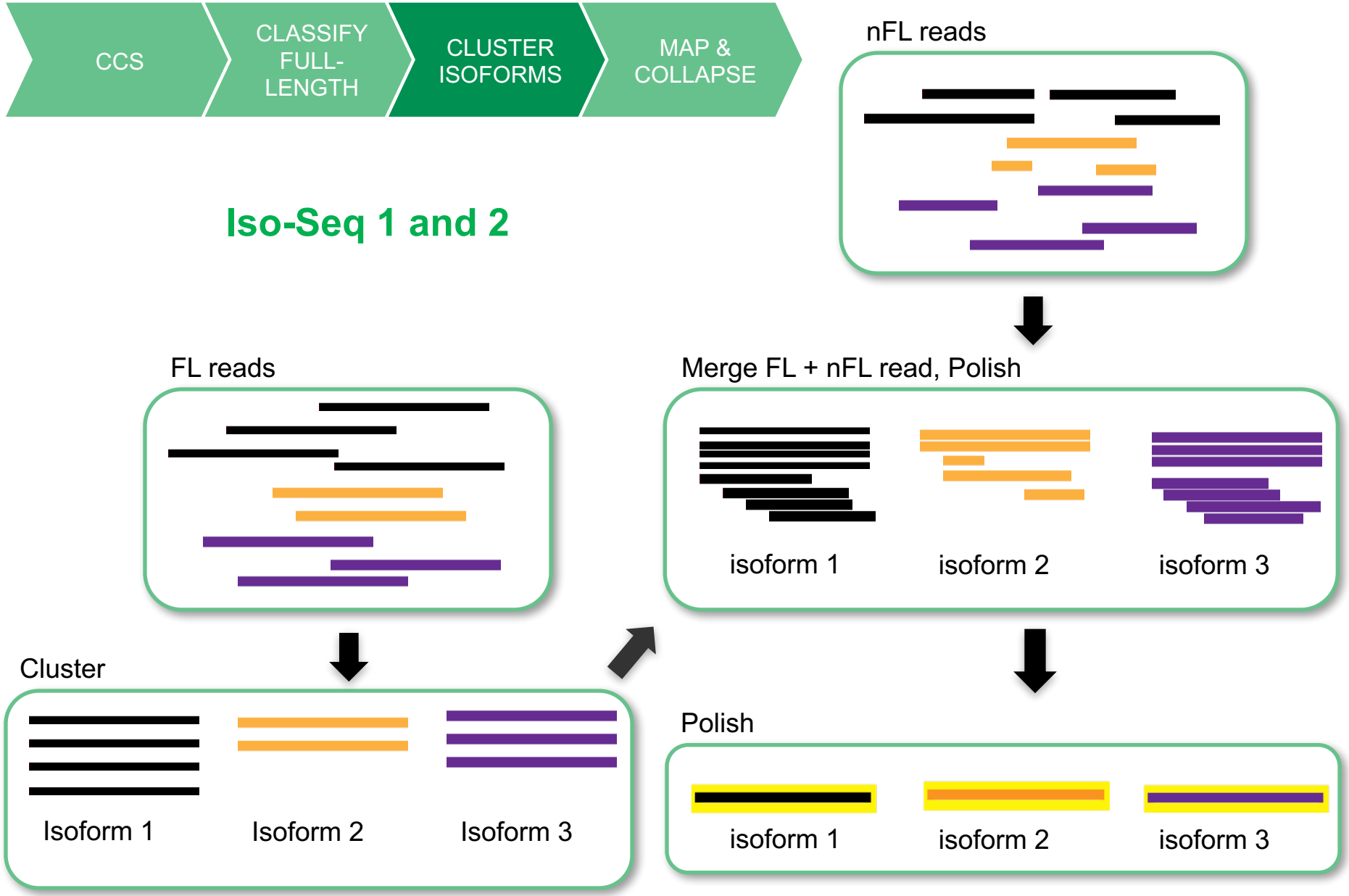
Iso-Seq1 & 2

FL reads

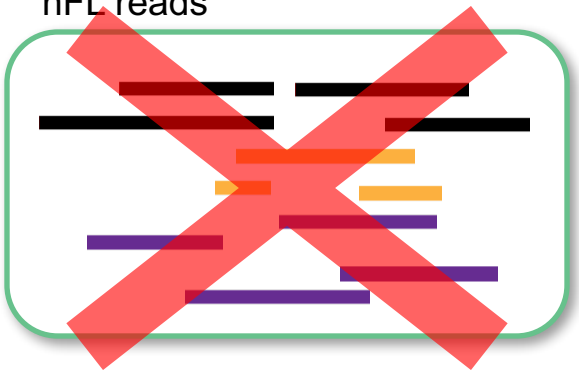
Full-length:

- Has 5’ cDNA primer
- Has 3’ cDNA primer
- Has polyA tail (>20 bp)

Support custom library prep and FL


CCS
CLASSIFY FULL-LENGTH
CLUSTER ISOFORMS
MAP & COLLAPSE
Iso-Seq 1 and 2
nFL reads
FL reads
Merge FL + nFL read, Polish
isoform 1
isoform 2
isoform 3
Cluster
Isoform 1
Isoform 2
Isoform 3
Polish
isoform 1
isoform 2
isoform 3

nFL reads

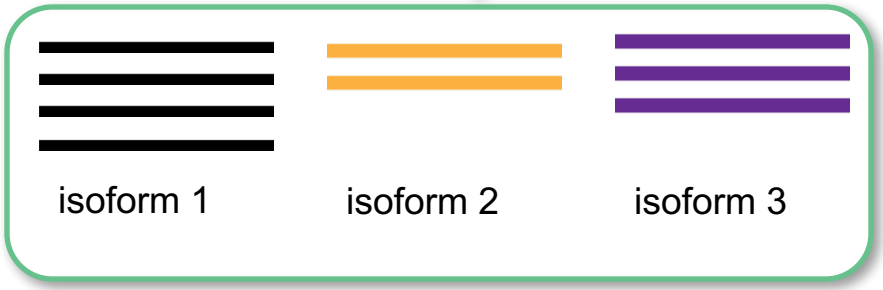


Iso-Seq 3

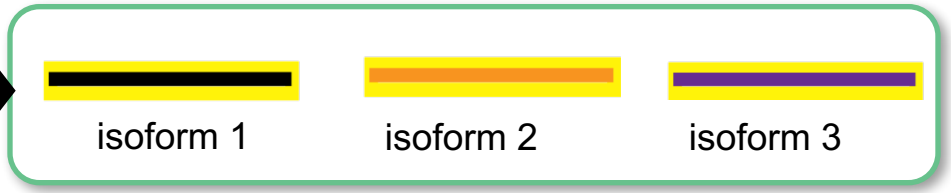
FL reads

Sequel higher throughput, longer reads, no nFL needed

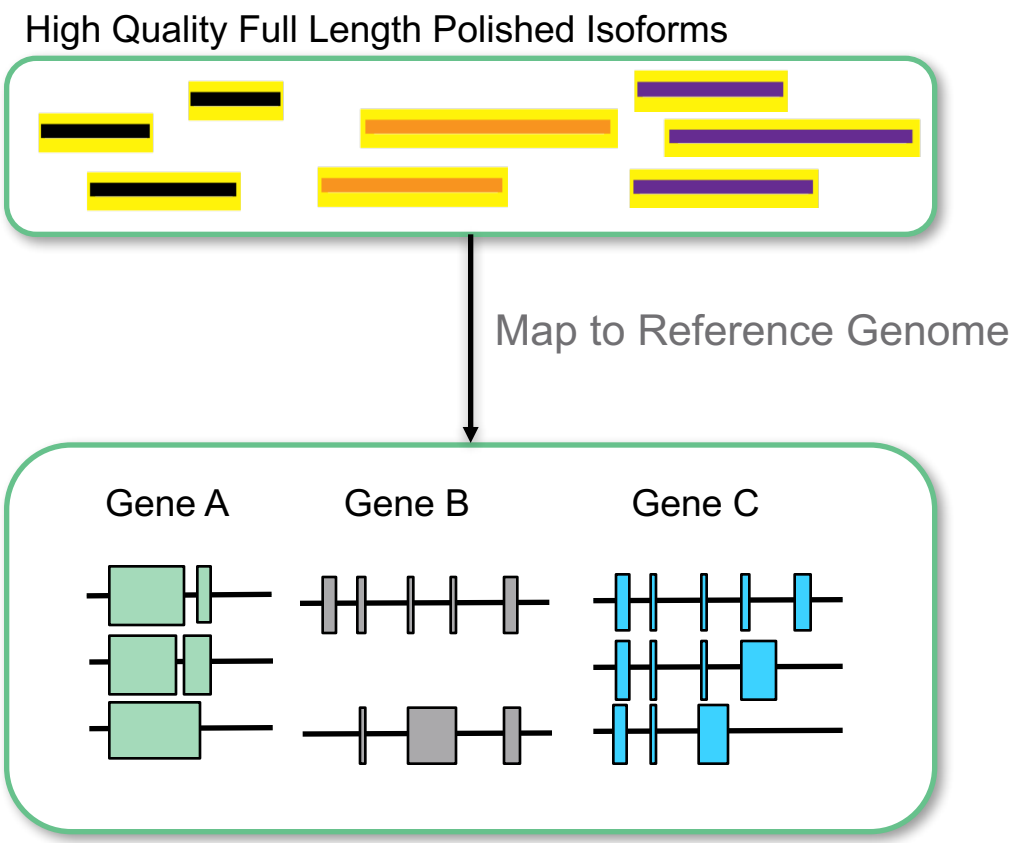
Cluster isoforms



Polish




CCS
CLASSIFY FULL-LENGTH CLASSIFY
CLUSTER ISOFORMS CLUSTER
MAP & COLLAPSE


High Quality Full Length Polished Isoforms
Map to Reference Genome
Gene A
Gene B
Gene C

SELECT YOUR ISO-SEQ WORKFLOW

Workflow	Output Results	Cases
Iso-Seq Classify Only	Full Length reads (FL) FASTQ	• Short amplicon (<1 kb) • Non-Eukaryotic（Bacteria, Virus）
Iso-Seq	Full Length High Quality Isoforms FASTQ	• No or poor Reference Genome • Eukaryotic
Iso-Seq w/ Mapping	Full Length High Quality, Collapsed Isoforms FASTQ, GFF	• Good Reference Genome • Eukaryotic

ISO-SEQ SUPPORTS MULTIPLEXING

Use Case: Same Species, Different Tissues/Timepoints

- Supported by SMRT Link
  - Use Iso-Seq analysis application in SMRT Link
  - Provide barcoded sequences as parameter to Classify step
- May use community script to get per barcode count information for each transcript after Iso-Seq is run

PACBIO®

Iso-Seq3

Ultra Fast + High Performance + Scalable

ISO-SEQ3 OVERVIEW

Iso-Seq3 workflow is the same as Iso-Seq1, 2

- CCS - same

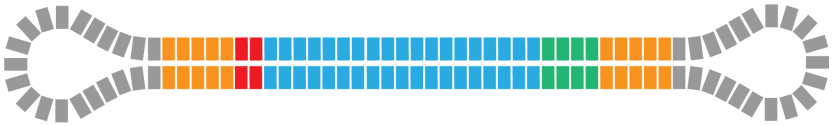
ISO-SEQ3 OVERVIEW

Iso-Seq3 workflow is the same as Iso-Seq1, 2 :

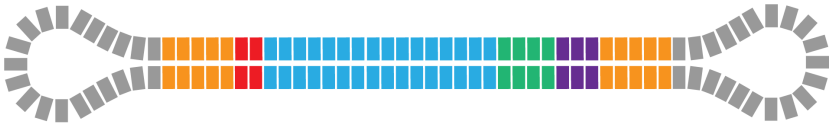
- CCS – same
- Classify – utilizing demultiplex barcoding algorithm (LIMA) with special `--isoseq` mode

ISO-SEQ3 CLASSIFY: LIBRARY PREP

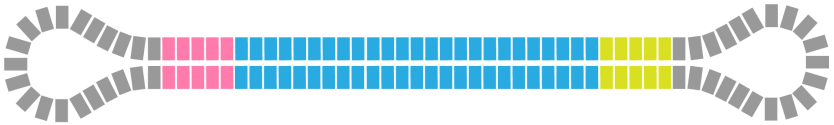
Whole transcriptome

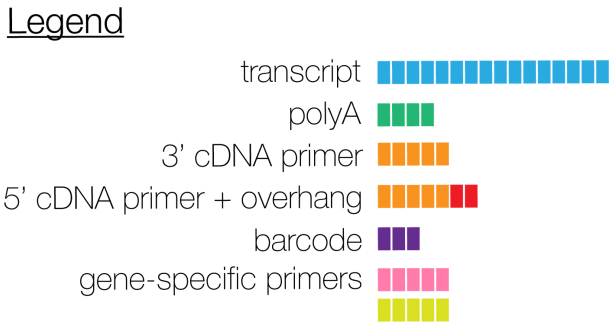


Whole transcriptome, barcoded



Targeted genes





ISO-SEQ3 CLASSIFY : ULTRA FAST

Sample	CCS READS	Iso-Seq3 Classify	Iso-Seq1 & 2 Classify
Maize, 1 cell	52,445	6 sec	40 min
Beef, 1 cell	110,067	15 sec	2 hr
Human, 10 cell	2,973,387	3 min	>8 hr+

- Iso-Seq3 Classify used less than 100 MB memory
- Test used 1 node * 16 CPU

ISO-SEQ3 CLASSIFY: MORE ACCURATE

Full Length:

TSO Artifact:

ISO-SEQ3 CLASSIFY: DETECT ARTIFICIAL CONCATEMER

Full Length:

TSO Artifact:

Artificial Concatemers:

- Due to insufficient SMRT adapters, fusion of two or multiple cDNA reads
- All Iso-Seq workflows remove concatemers

LIBRARY ARTIFACTS

Type	Cause	ISO-SEQ1 & 2 can detect	ISO-SEQ3 can detect
TSO Artifacts	Template switching artifacts	no	yes
Artificial Concatemers	Insufficient SMRT adapter	yes	yes

ISO-SEQ3 WORKFLOW

Iso-Seq3 workflow is the same as Iso-Seq1 & 2

- CCS - same
- Classify – utilizing demultiplex barcoding algorithm (LIMA) with special `--isoseq` mode
- Cluster - faster, better results

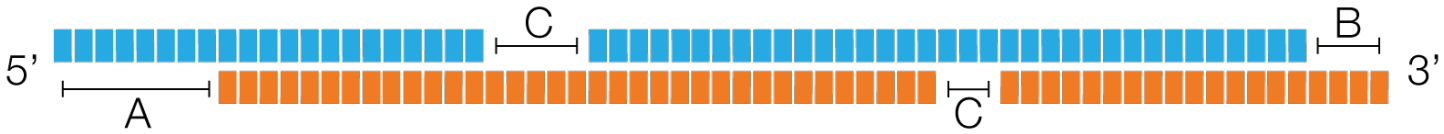
ISO-SEQ3 CLUSTER: ISOFORM DEFINITION

Two Full-Length reads are considered ‘similar’ if they are:

(A) <100 bp difference in 5’ start

(B) <30 bp difference in 3’ end

(C) <10 bp in internal gap (exon)



ISO-SEQ3 : POLISH ISOFORMS

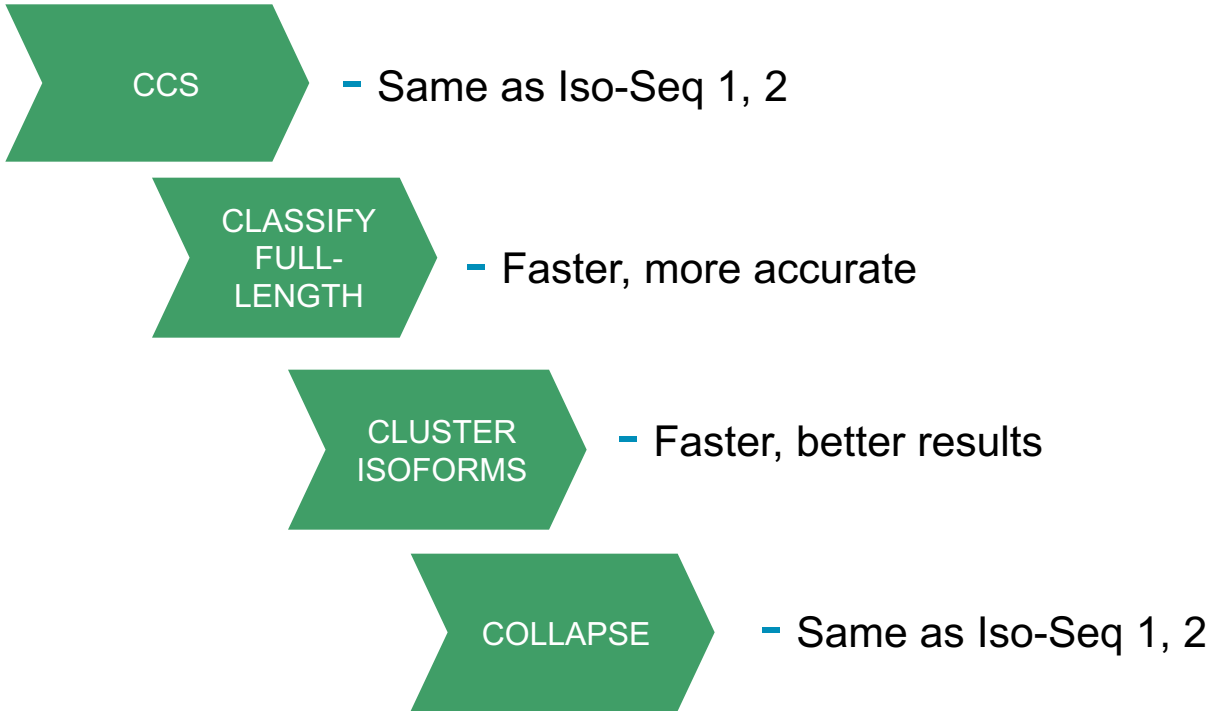
The Polish step generates consensus sequences which are divided into:

- High Quality (HQ): accuracy ≥99% AND ≥2 FL read support
- Low Quality (LQ): accuracy <99% +

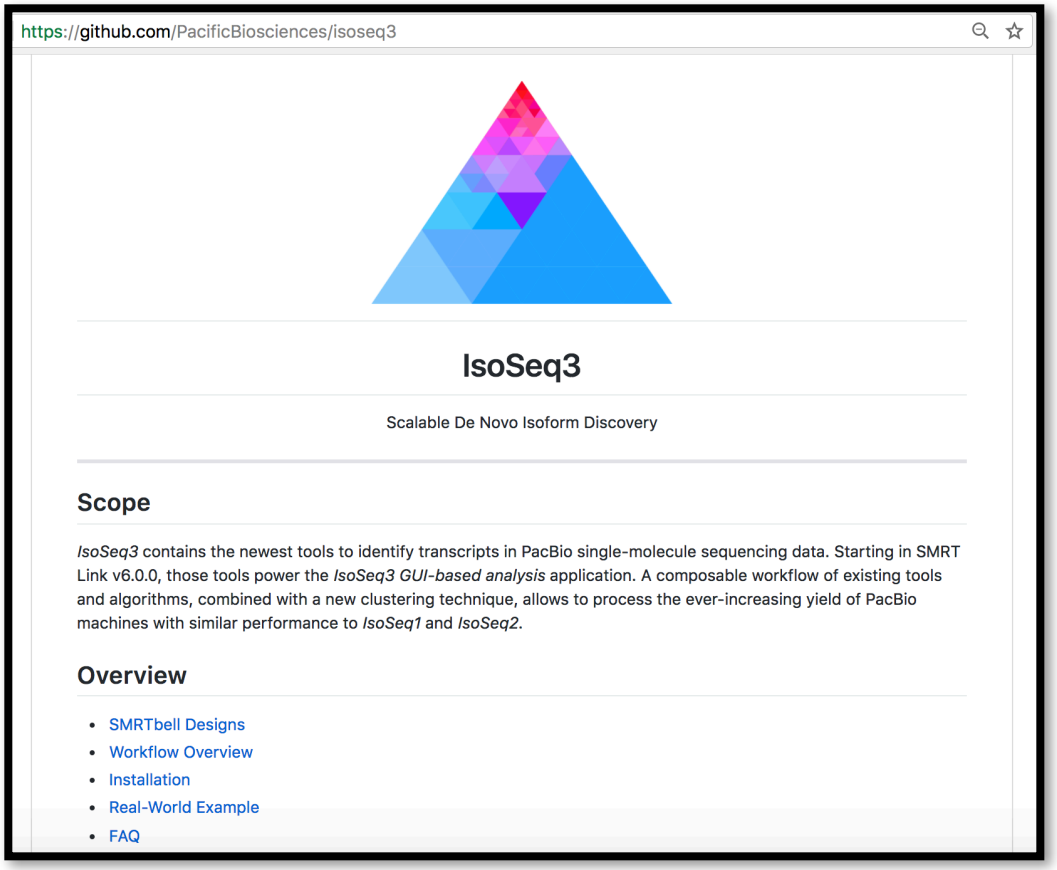
Recommend to only look at HQ isoforms

+ In Iso-Seq3, unclustered (singleton) FL reads are not output. Both HQ/LQ are supported by 2 or more FL reads and only differentiated by predicted accuracy.

ISO-SEQ3 IMPROVEMENT



- Written in C++, faster, less memory, better results



IsoSeq3 GitHub stand alone binary for advanced users, NO official Tech Support

Report bugs to GitHub Issues

Official release in SMRT Link v6.0

PUBLIC 1 CELL SEQUEL DATA

Download Link: https://downloads.pacbcloud.com/public/dataset/RC0_1cell_2017

Secure https://downloads.pacbcloud.com/public/dataset/RC0_1cell_2017/

Index of /public/dataset/RC0_1cell_2017

Name	Last modified	Size	Description
Parent Directory		-	
README.txt	2017-08-08 13:47	2.0K	
isoseq_flnc.fasta	2017-06-17 22:53	496M	
isoseq_nfl.fasta	2017-06-17 22:53	248M	
m54086_170204_081430.adapters.fasta	2017-02-04 11:42	58	
m54086_170204_081430.scraps.bam	2017-02-04 11:41	12G	
m54086_170204_081430.scraps.bam.pbi	2017-02-04 11:41	35M	
m54086_170204_081430.sts.xml	2017-02-04 11:41	96K	
m54086_170204_081430.subreads.bam	2017-02-04 11:39	8.8G	
m54086_170204_081430.subreads.bam.pbi	2017-02-04 11:39	23M	
m54086_170204_081430.subreadset.xml	2017-02-04 11:37	10K	

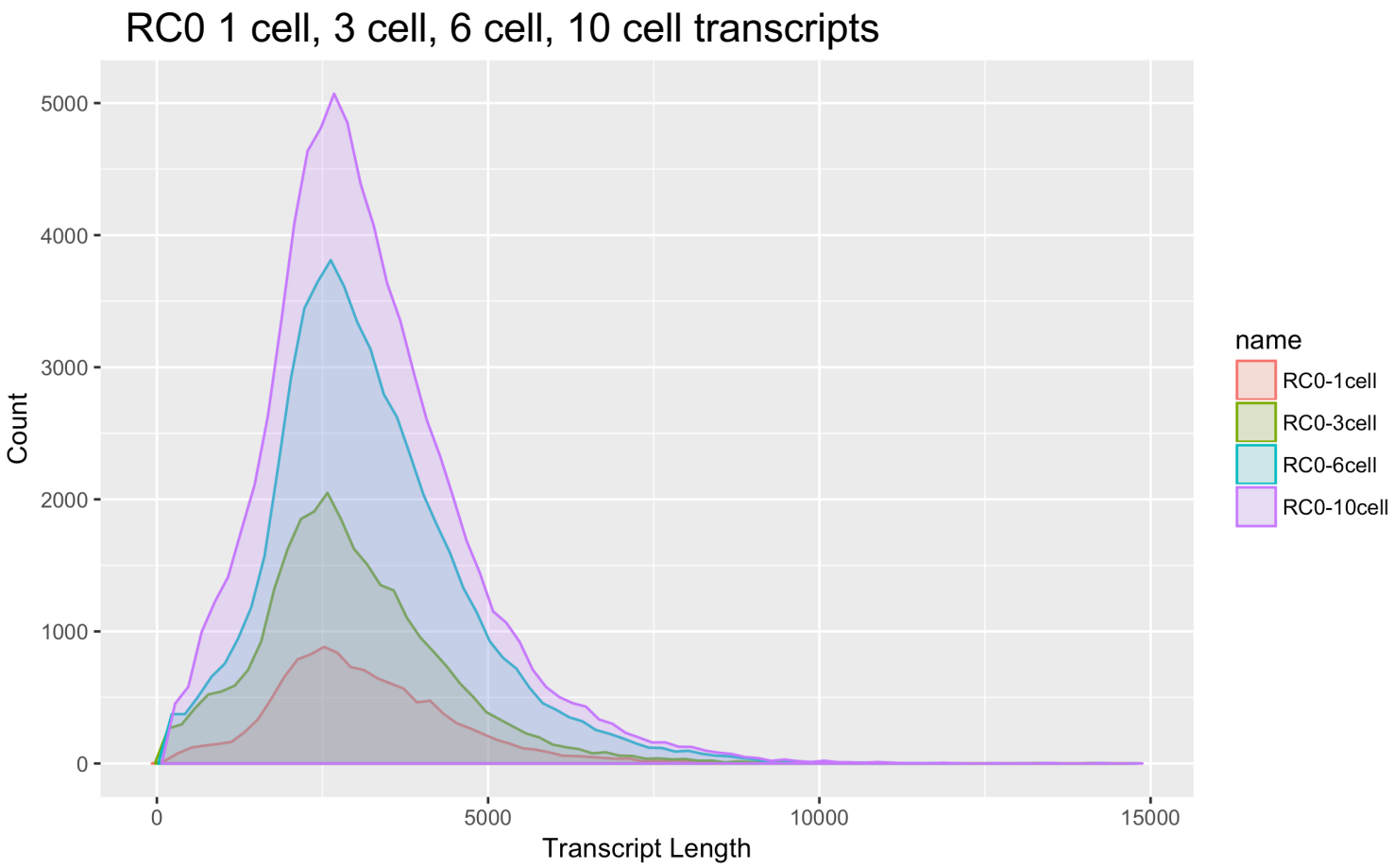
Iso-Seq3 Performance

ISO-SEQ3 IS FAST

SAMPLE	SMRT CELLS	FL READS	CLASSIFY	CLUSTER	POLISH
RC0	1	182,211	19 sec	8 min	2.5 hr
RC0	3	568,541	1 min	21 min	11 hr
RC0	6	1,327,856	2 min	1 hr	3 hr per node (24 nodes)
RC0	10	2,038,060	3 min	2 hr	3 hr per node (24 nodes)
Mouse Liver	2	259,081	13 sec	4 min	4 hr

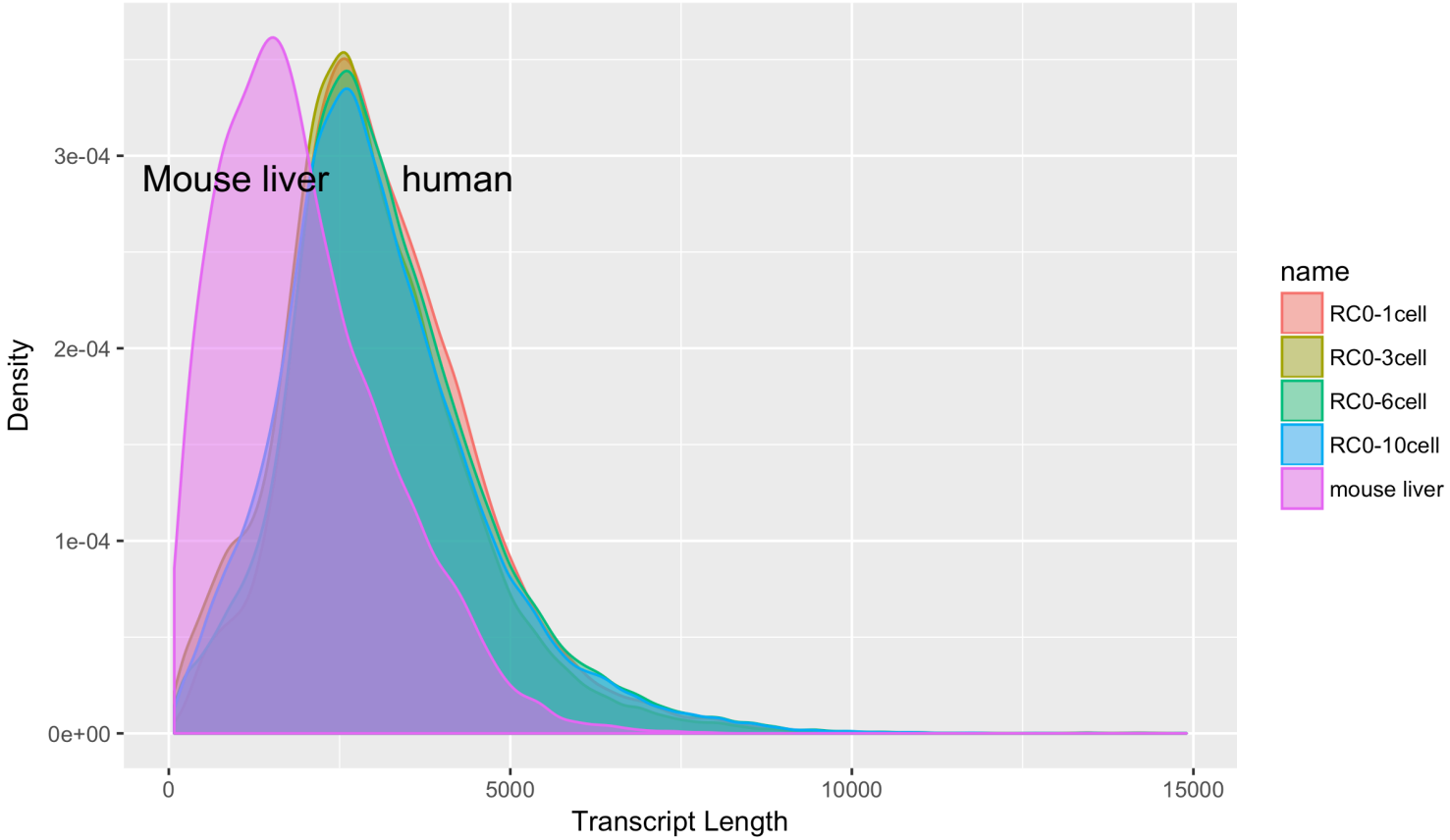
- RC0 = Universal Human Reference RNA (human) + Lexogen SIRV spike-in controls
- Not including CCS and Mapping runtime
- Computing configuration：16 CPU / node
- Tested using command line

HUMAN TRANSCRIPTS LENGTH DISTRIBUTION



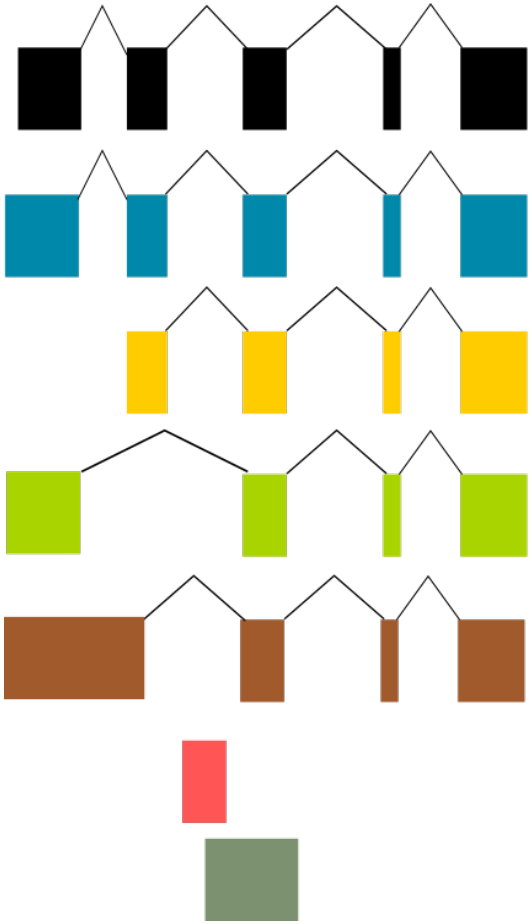
DIFFERENCE BETWEEN HUMAN AND MOUSE LIVER TRANSCRIPTS

Mouse liver transcripts slightly shorter than RC0



USE SQANTI* TO EVALUATE ISO-SEQ3 RESULTS

*SQANTI is a community tool developed by Conesa lab



Reference

FSM Full Splice Match, matches reference perfectly.

ISM Incomplete Splice Matches, matches reference partially

NIC Novel In Catalog, novel isoform using known junctions

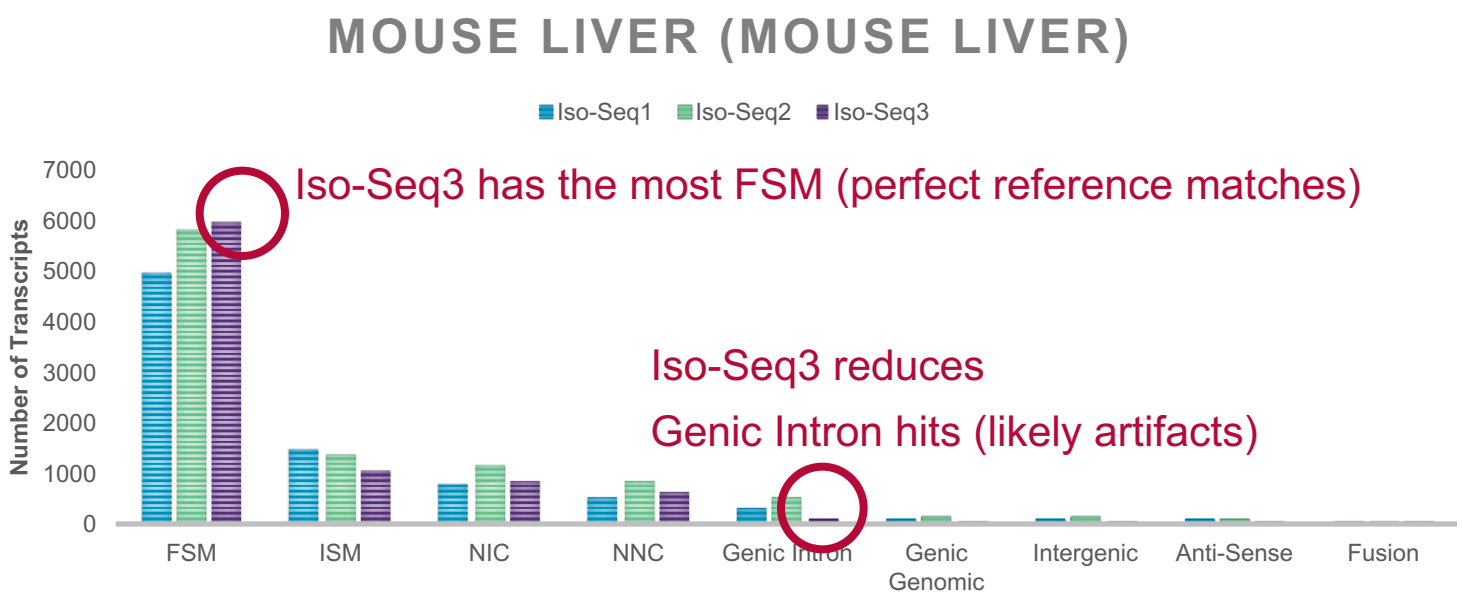
NNC Novel Not in Catalog, novel isoforms using novel junctions.

Genic Intron within intron

Genic Genomic Overlap with intron and exons

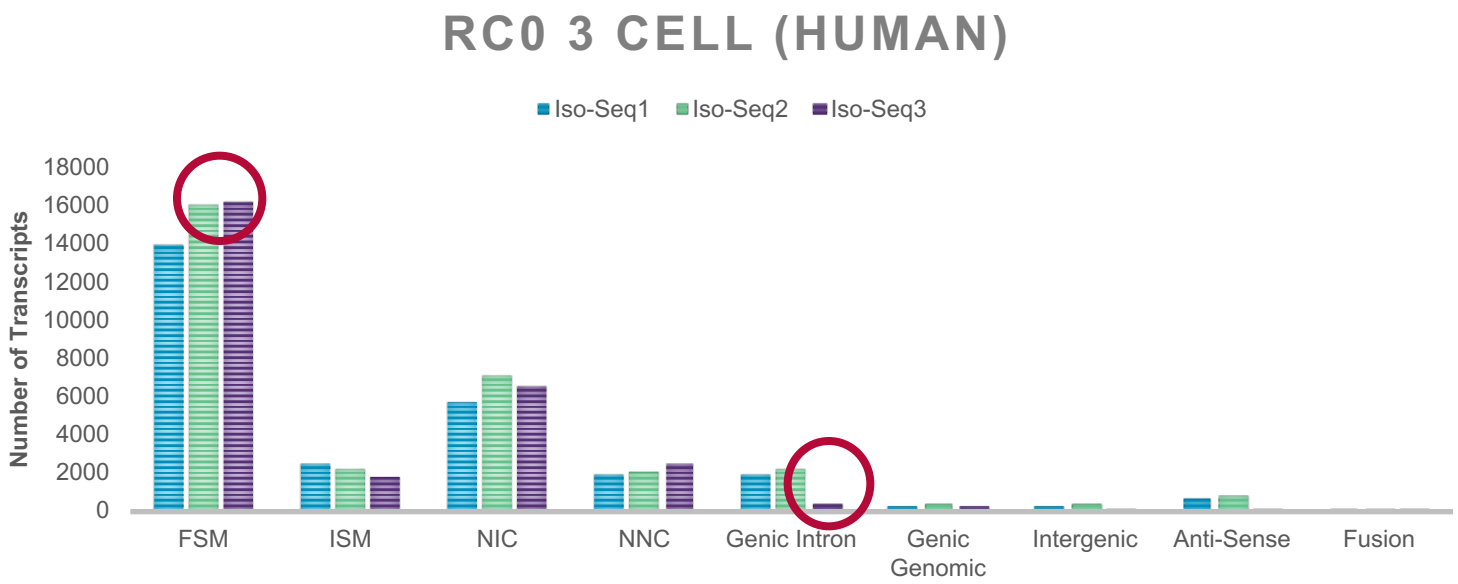
Tardaguila, M. *et al.* SQANTI: extensive characterization of long read transcript sequences for quality control in full-length transcriptome identification and quantification. 1–31 (2017). doi:10.1101/118083

ISO-SEQ3 VS REF ANNOTATION: MOUSE LIVER



SQANTI : compare Iso-Seq results vs Gencode M16 Reference Gene Annotation

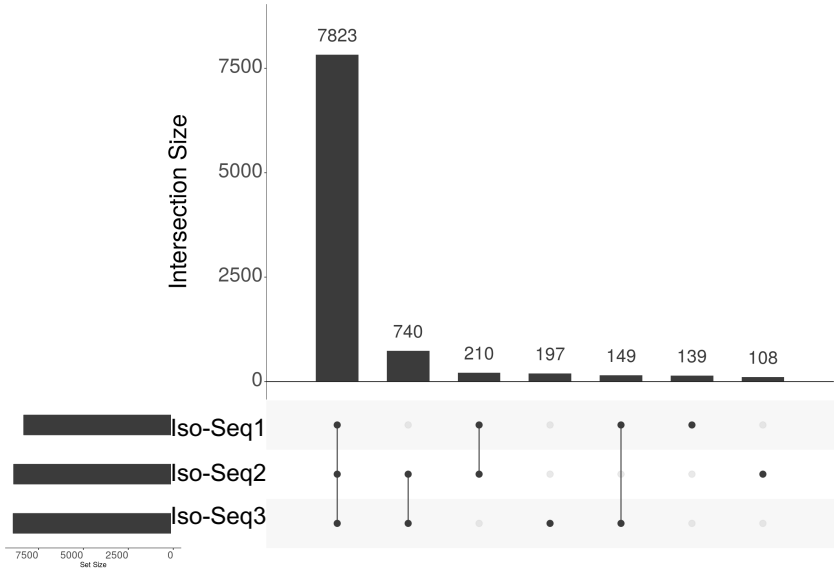
ISO-SEQ3 VS REF ANNOTATION: HUMAN



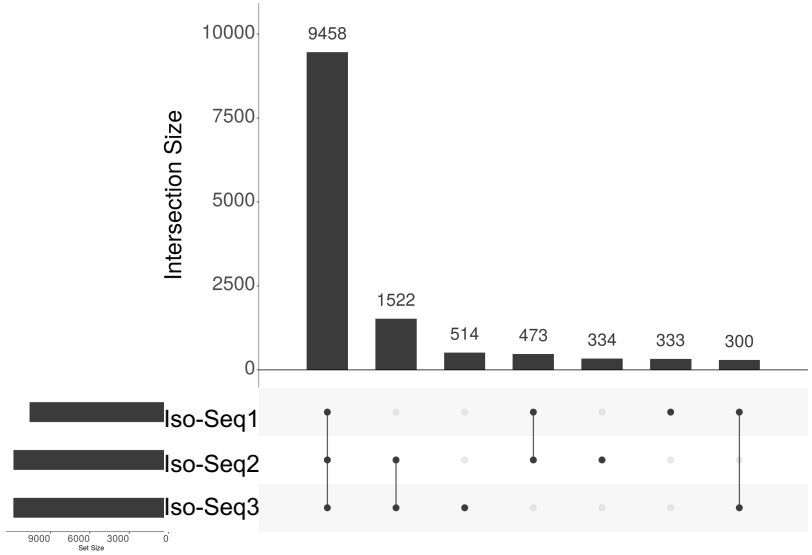
SQANTI : compare Iso-Seq results vs Gencode v27 Reference Gene Annotation

ISO-SEQ (1, 2, 3) GENERATE CONSISTENT RESULTS

RC0 3 Cells, Known Genes Only



RC0 3 Cells, Known Isoforms Only



* Only report FSM gene and isoforms

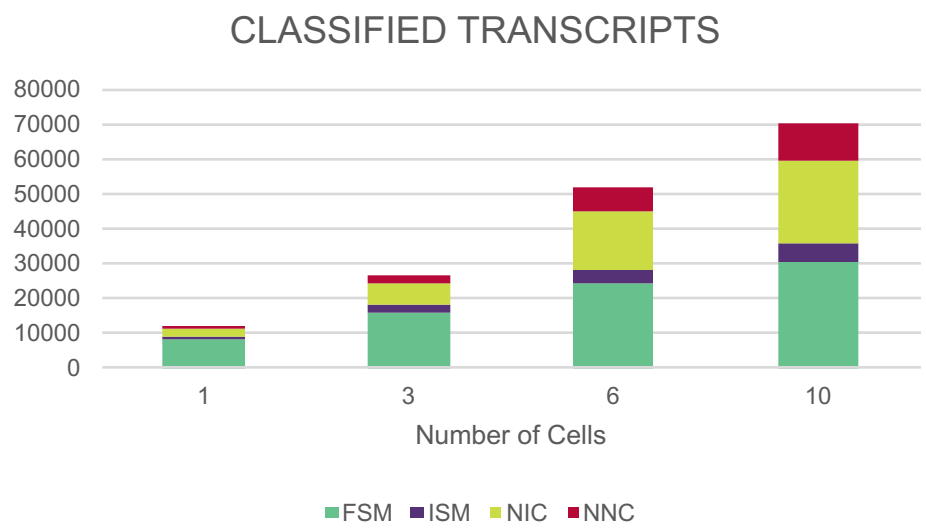
HOW MUCH SEQUENCING IS NEEDED?

FSM = Full Splice Match

ISM = Incomplete Splice Matches

NIC = Novel In Catalog

NNC = Novel Not in Catalog



PACBIO®

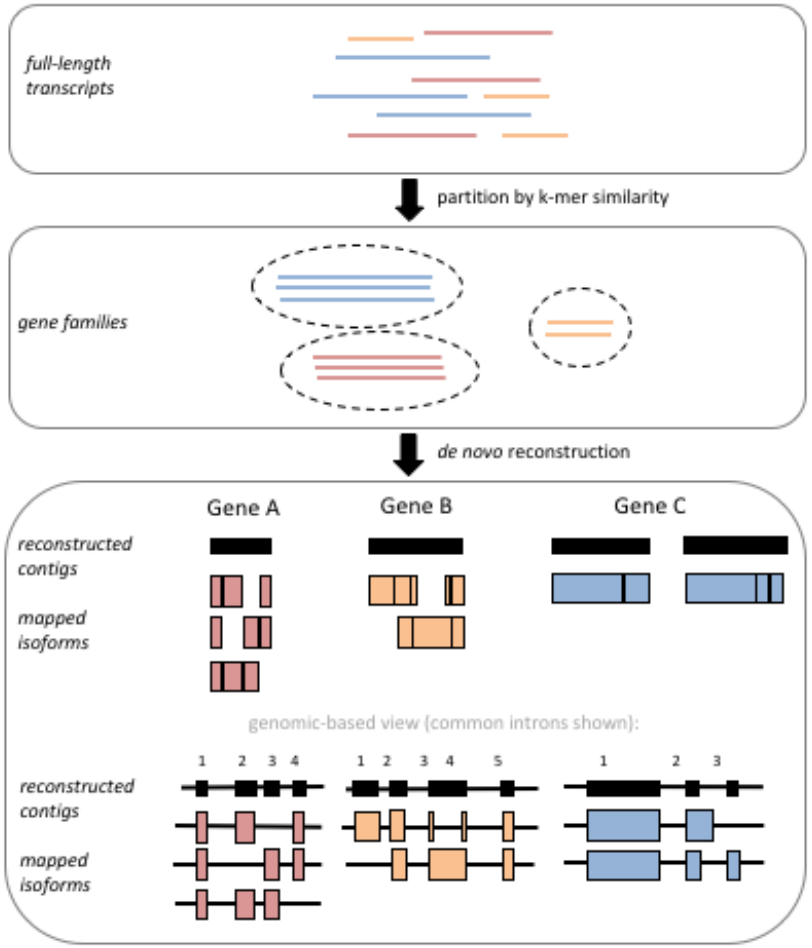
Iso-Seq Community Tools

COGENT: RECONSTRUCT CODING REGION

- Cogent
- No or poor reference genome

- Input: Iso-Seq high-quality isoforms
- Outpu: reconstructed coding regions

- Reconstructed coding regions can be used to:
  - Collapse isoforms
  - Infer gene count
  - Evaluate genome assemblies



CUPCAKE & TAMA: LIGHT-WEIGHT ANALYSIS SCRIPTS

Cupcake has many Iso-Seq downstream analysis scripts

- Remove redundant isoforms
- Merge Iso-Seq runs from different batches
- Junctions analysis
- Estimate probe enrichment on-target rate
- Plot rarefaction curve: infer sequencing coverage and gene count

TAMA, developed by PacBio user Richard Kuo

- Remove redundant isoforms
- Merge Iso-Seq runs from different batches
- Predict ORF, and Nonsense Mediated Decay (NMD)

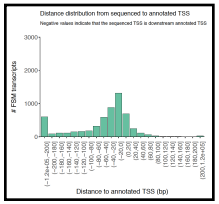
Iso-Seq Community Tools List

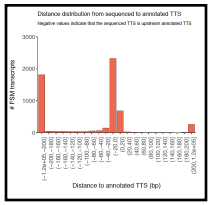
SQANTI & TAPPAS: QUALITY CONTROL, EVALUATION AND VISUALIZATION

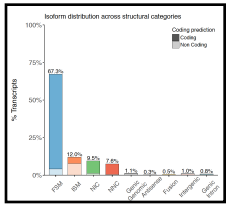
Developed by Ana Conesa Lab (U of FL)

SQANTI

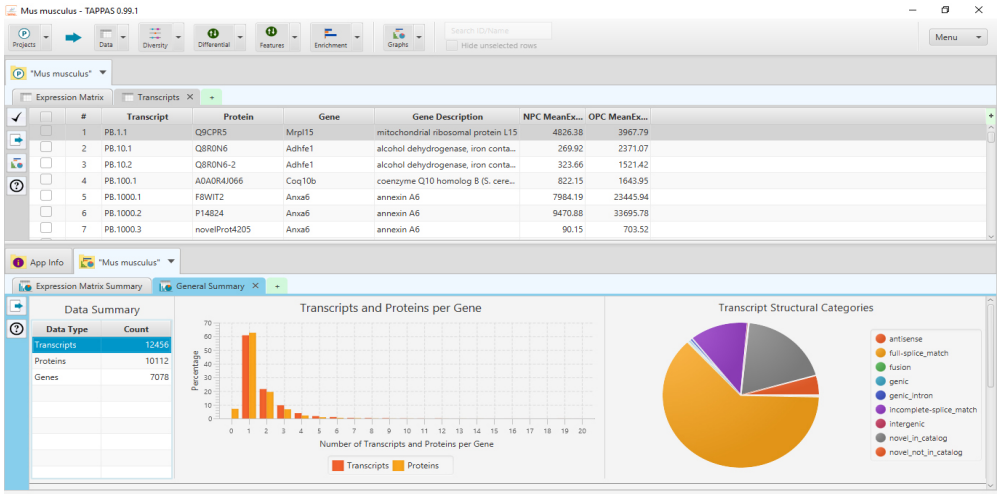
- Compare with annotation
- Detect and remove artifacts
- Combine with RNA-seq data
- Output PDF report







TAPPAS visualize data at isoform level



Google Group:

groups.google.com/forum/#!forum/SMRT_isoseq

GitHub Repository and Tutorials:

github.com/PacificBiosciences/IsoSeq_SA3nUP/
(**http://tinyurl.com/PBisoseq)**

https://github.com/PacificBiosciences/IsoSeq3

PACBIO®

www.pacb.com