SMRT Leiden Iso-Seq Session: Tools, Tools, Tools

Elizabeth Tseng, PacBio

@Magdoll

@Magdoll

Slides will be posted on Twitter and Google Group.

Online Resources:

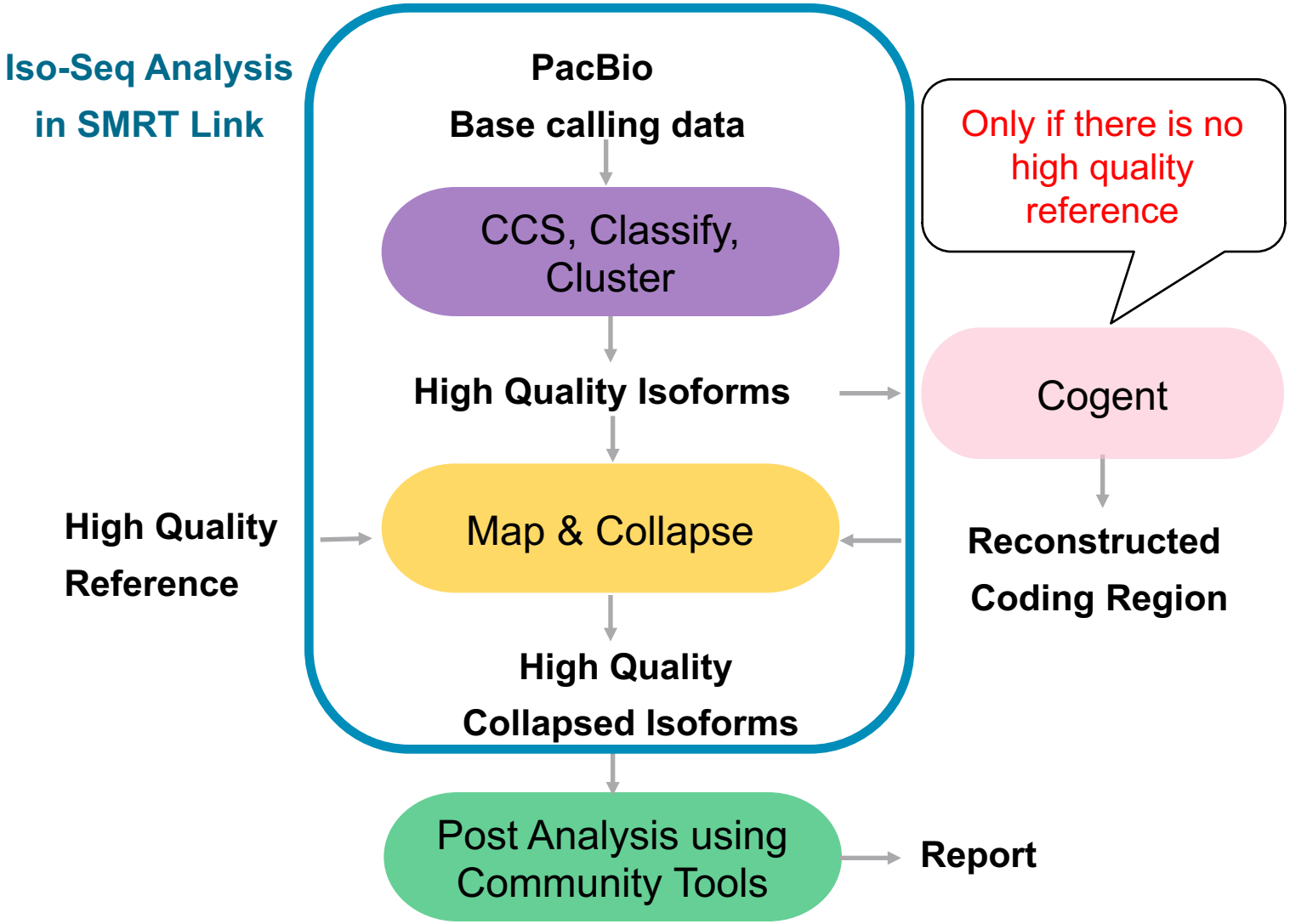
groups.google.com/forum/#!forum/SMRT_isoseq

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

OVERVIEW

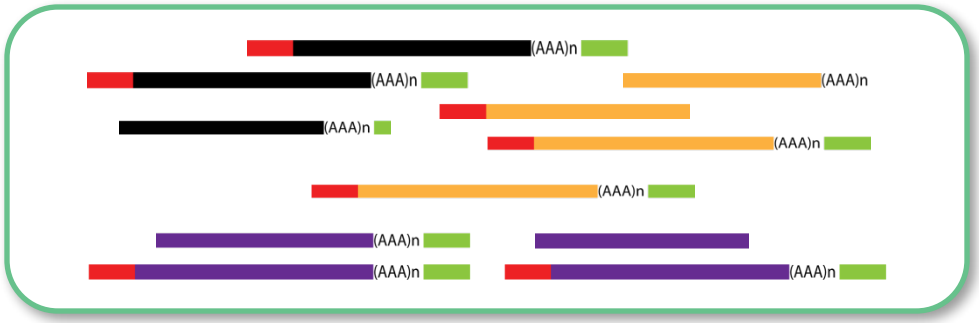
- Recommended Iso-Seq Bioinformatics Workflow
- Developers Version of Iso-Seq3
- List of Iso-Seq community tools
- Aligners: GMAP, minimap2, or…?

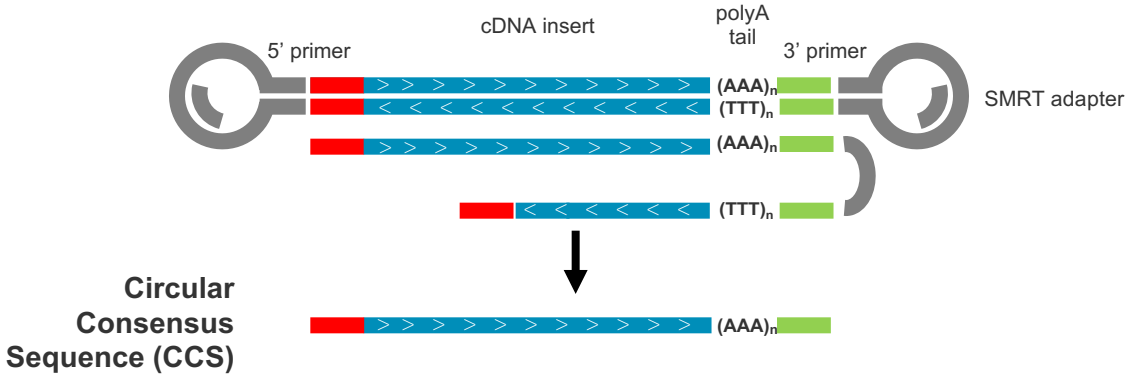
ISO-SEQ ANALYSIS WORKFLOW



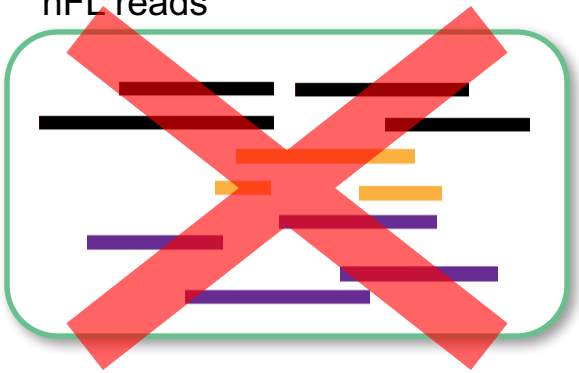

CCS
CLASSIFY FULL-LENGTH
CLUSTER ISOFORMS CLUSTER
MAP & COLLAPSE

CCS


(AAA)n


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

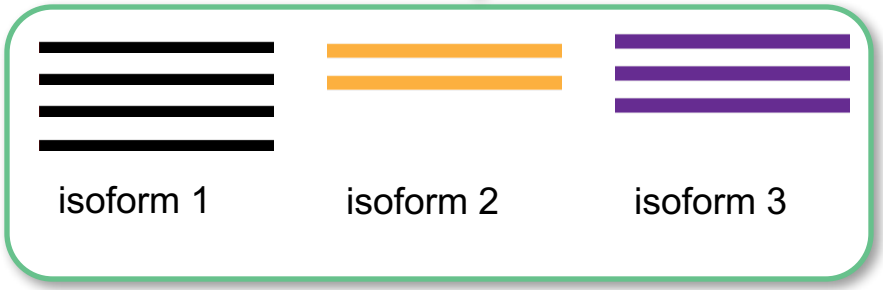
nFL reads



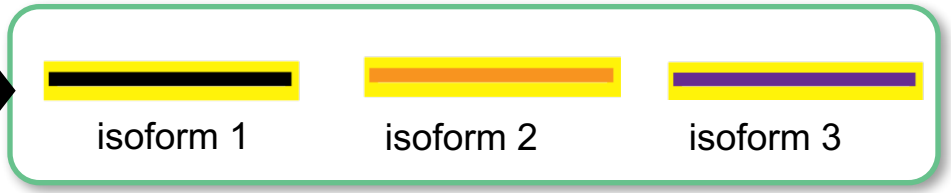
Iso-Seq3

FL reads

Cluster isoforms

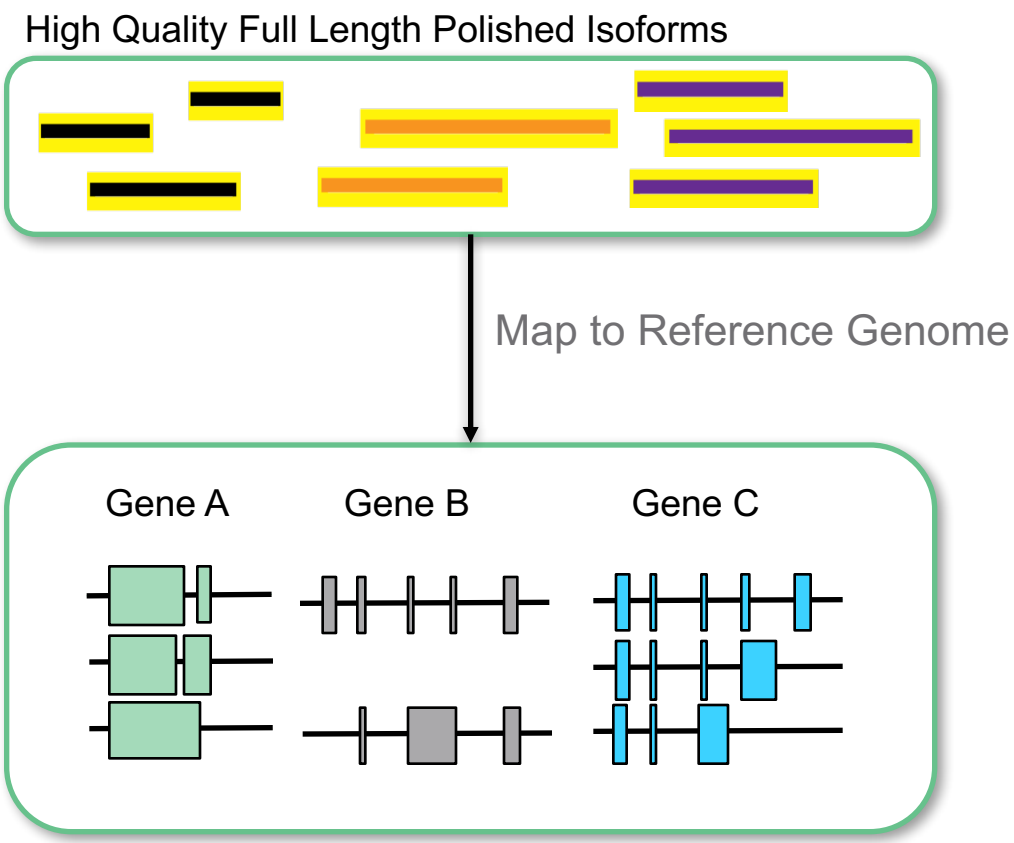


Polish



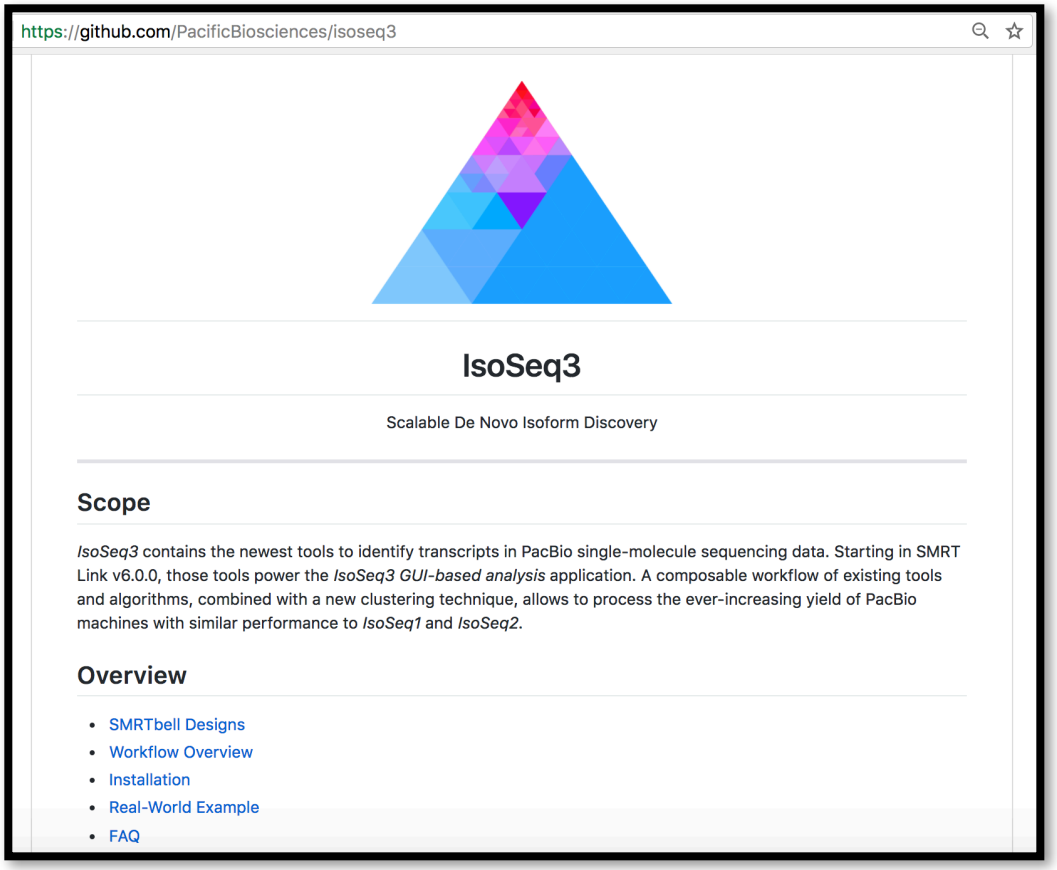
Sequel System has higher throughput, longer reads, no nFL needed


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

OVERVIEW

- Recommended Iso-Seq Bioinformatics Workflow
- Developers Version of Iso-Seq3
- List of Iso-Seq community tools
- Aligners: GMAP, minimap2, or…?



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 SYSTEM 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	

OVERVIEW

- Recommended Iso-Seq Bioinformatics Workflow
- Developers Version of Iso-Seq3
- List of Iso-Seq community tools
- Aligners: GMAP, minimap2, or…?

ISO-SEQ BIOINFX: NEEDS AND SOLUTIONS (1)

Error Correction

Goal: achieve sufficient accuracy to perform downstream analysis (99-100%)
Methods: genome-guided or *de novo*; PacBio-only or hybrid
Challenge: scalability, indel errors
Tools: Iso-Seq, IsoCon, LSC+IDP, IDP-denovo

Spliced Aligner

Goal: align to genome to perform downstream analysis
Challenge: scalability, indel errors affecting junction mapping
Tools: GMAP, STAR, minimap2

Alignment Processing Tools

Goal: Alignment filtering, collapsing redundant or degraded transcripts, etc
Tools: Cupcake, TAMA

Comparative Tools / Annotation Tools

Goal: identify novel isoforms/genes against reference annotation
Tools: matchAnnot, SQANTI, CAT, LoReAn

ISO-SEQ BIOINFX: NEEDS AND SOLUTIONS (2)

ORF Prediction

Goal: Open Reading Frame prediction that is robust to errors

Challenge: scalability, indel errors

Tools: ANGEL

lncRNA Prediction

Tools: lncRNA pipeline

Data Visualization and Protein/Isoform Analysis

Tools: TAPPAS

Coding Genome Reconstruction without a Genome

Goal: Reconstruct the coding portions of gene loci using Iso-Seq data only

Tools: Cogent

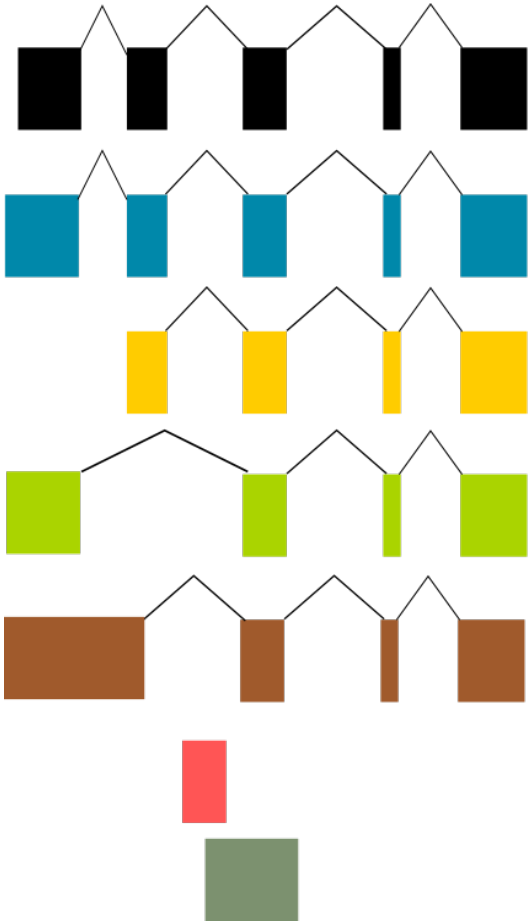
Phasing / Allele Specific Expression

Goal: Phasing diploid or tetraploid Iso-Seq data

Tools: IsoPhase (PAG2018 presentation here)

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

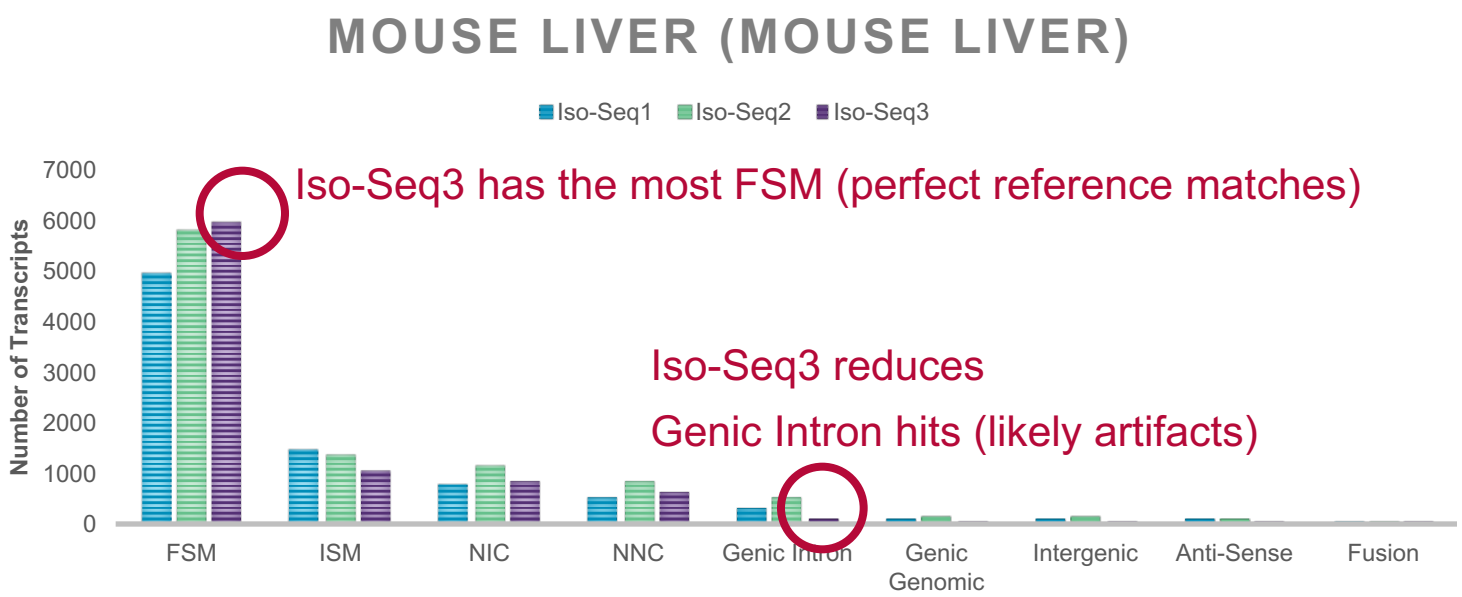
NNC Novel Not in Catalog, novel isoforms using novel junctions.

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

OVERVIEW

- Recommended Iso-Seq Bioinformatics Workflow
- Developers Version of Iso-Seq3
- List of Iso-Seq community tools
- Aligners: GMAP, minimap2, or…?

DATASETS

Species	Human	SIRV	Maize
Number of Sequences	92,924	92,924	280,473

The human and SIRV dataset comes from the same 6-cell RC0 (human + SIRV) run. The sequences will not be pre-separated so the “unmapped read” count will be high for the SIRV (since most of RC0 is human).

The maize dataset comes from Wang et al. (2016), a six-tissue Iso-Seq dataset of maize.

The input from all three are “HQ isoform sequences”. That is, the output from Iso-Seq clustering which is *de novo* (no ref genome or annotation guided).

PARAMETER

GMAP parameter:

```
gmap -n 0 -t 30 -z sense_force –-cross-species --max-intronlength-ends 200000
```

Minimap2 parameter:

```
minimap2 -ax splice -uf --secondary=none -t 30 -C5
```

- gmap version 2018-03-20 vs 2018-05-30
- minimap2 version 2.9-r720
- gmap DB and minimap2 .mmi provided
- Use hg38_noalt --- hg38 NOT including alt contigs! (fasta provided by Heng Li)
- Gencode v27 annotation (renamed, provided by Heng Li)

RUNTIME

	gmap-0320	gmap-0530	minimap2
Human	19 min	60 min	2 min
SIRV	12 sec	10 sec	4 sec
Maize	Crashed	16 min	1 min

- GMAP index build time for human: 25 min

- For maize genome and annotation, we intentionally choose a version that pre-dates the inclusion of the same 6-tissue Iso-Seq data incorporation. The latest maize B73 annotation (v4) uses Iso-Seq. We choose a 2015 release (v3.22) that did not include Iso-Seq data.
  - genome: Zea_mays.AGPv3.22.dna_rm.genome.fa
  - annotation: release 5b+

MAPPABILITY

HUMAN	gmap-0320	gmap-0530	minimap2
Input	92,924	92,924	92,924
Unmapped	392	392	402
Mapped Chimeric	568	561	977
Mapped Non-Chimeric	**91,964**	91,545	91,545

SIRV	gmap-0320	gmap-0530	minimap2
Input	92,924	92,924	92,924
Unmapped	92,684	92,684	92,684
Mapped Chimeric	5	5	9
Mapped Non-Chimeric	**235**	231	231

MAIZE	gmap-0320	gmap-0530	minimap2
Input	280,473	280,473	280,473
Unmapped	CRASHED	5020	2687
Mapped Chimeric	CRASHED	7356	9364
Mapped Non-Chimeric	CRASHED	268,097	**268,422**

SPLICE JUNCTIONS

HUMAN	gmap-0320	gmap-0530	minimap2
Total Junctions	780,419	780,738	775,805
Canonical Junctions	758,950	758,794	757,560
Junctions matching annotation	**758,314**	758,247	756,770
Junctions within 5 bp of annotation	759,198	**759,370**	757,944

SIRV	gmap-0320	gmap-0530	minimap2
Total Junctions	969	969	968
Canonical Junctions	923	922	920
Junctions matching annotation	**946**	945	942
Junctions within 5 bp of annotation	**946**	**946**	943

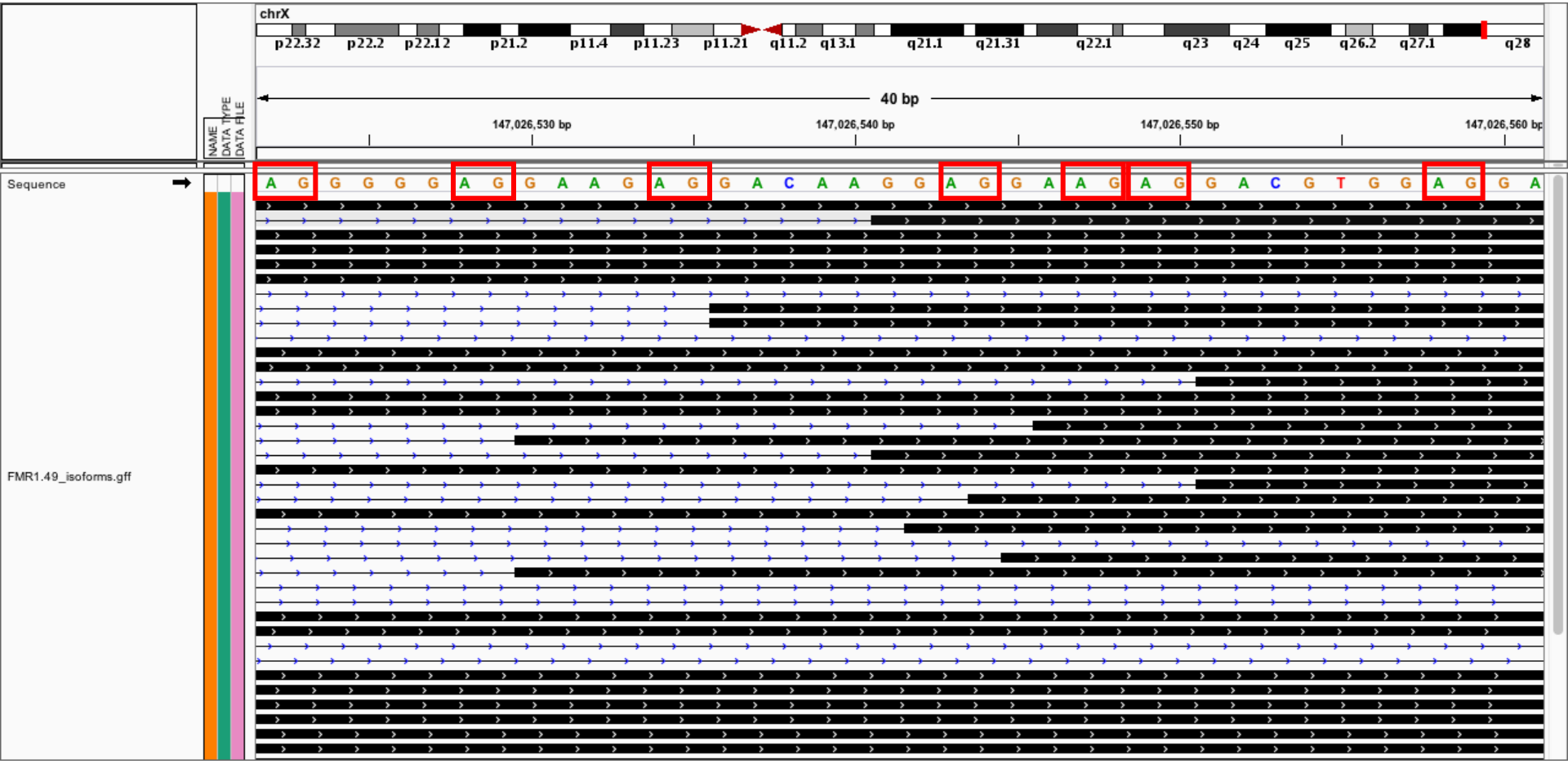
MAIZE	gmap-0320	gmap-0530	minimap2
Total Junctions	CRASHED	1,119,505	1,159,543
Canonical Junctions	CRASHED	1,045,593	1,052,350
Junctions matching annotation	CRASHED	**874,411**	872,958
Junctions within 5 bp of annotation	CRASHED	**880,093**	876,975

PRECISE JUNCTION MAPPING IS IMPORTANT

In this case, there's sufficient reads supporting the RNA editing and the alt junction.

What if there were fewer reads? What if the alt junction is only 1 bp from the canonical?

SOMETIMES, NATURE LIKE TO PUT A LOT OF CANDIDATE SPLICING ACCEPTOR SITES TOGETHER!



ALIGNER COMPARISON SUMMARY

- GMAP is slower than minimap2 by orders of magnitude
- However, GMAP seems to be slightly better at aligning precise junctions
- Different versions of GMAP perform differently!

- Both aligners are continuously evolving

Advice:

1. Always use the latest version of the aligners
2. Run both aligners and compare

PACBIO®

www.pacb.com