The Why, What, and How of the Iso-Seq Method:

Using Full-length RNA Sequencing to Annotate Genomes and Solve Diseases

PACBIO®

Intro to the Iso-Seq Method

Emily Hatas, senior director of business development

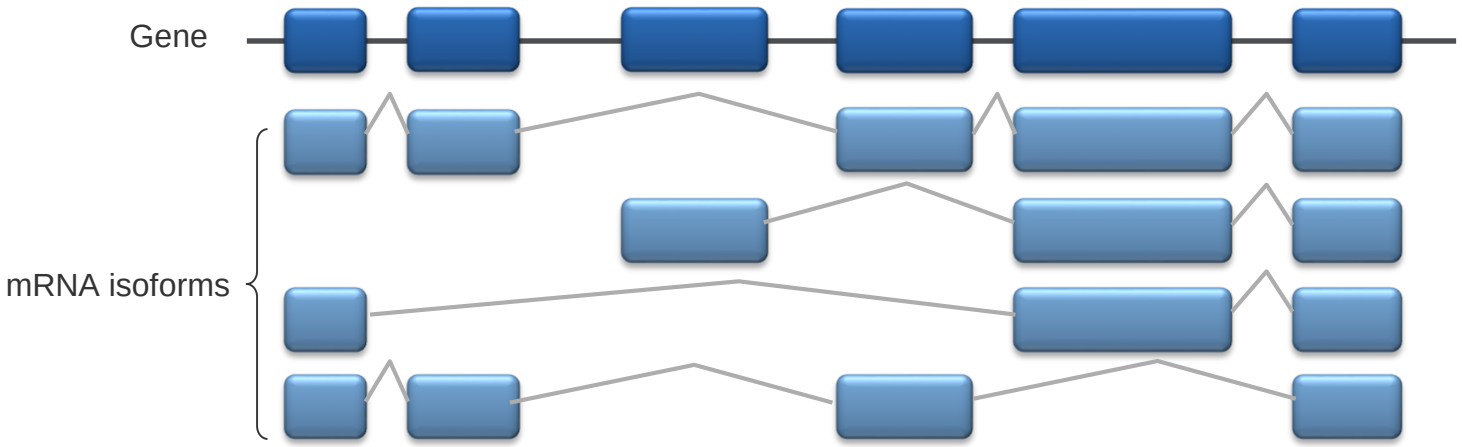
@EmilyHatas

WHAT IS THE ISO-SEQ METHOD?

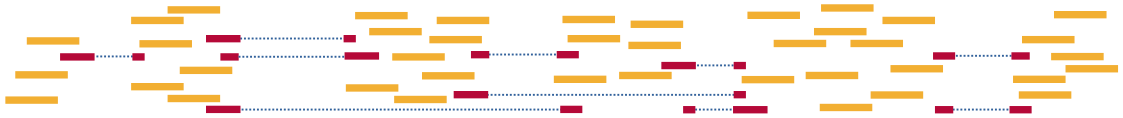
The PacBio Iso-Seq method is an end-to-end workflow for sequencing and analyzing full-length transcript isoforms.

1. Convert RNA → cDNA
2. cDNA → SMRTbell library
3. Sequence on the Sequel System
4. Generate circular consensus sequences (CCS)
5. Discover isoforms *de novo* with Iso-Seq analysis

WHY IS FULL-LENGTH RNA SEQUENCING USEFUL?



Short-read RNA-seq:



Reads spanning splice junctions

Insufficient Connectivity
Splice Isoform Uncertainty

Iso-Seq method:

Full-length cDNA Sequence Reads

Splice Isoform Certainty – No Assembly Required

KEY APPLICATIONS OF THE ISO-SEQ METHOD

Whole-genome Annotation

"I would like a reference catalog of all transcript isoforms detectable within a particular sample."

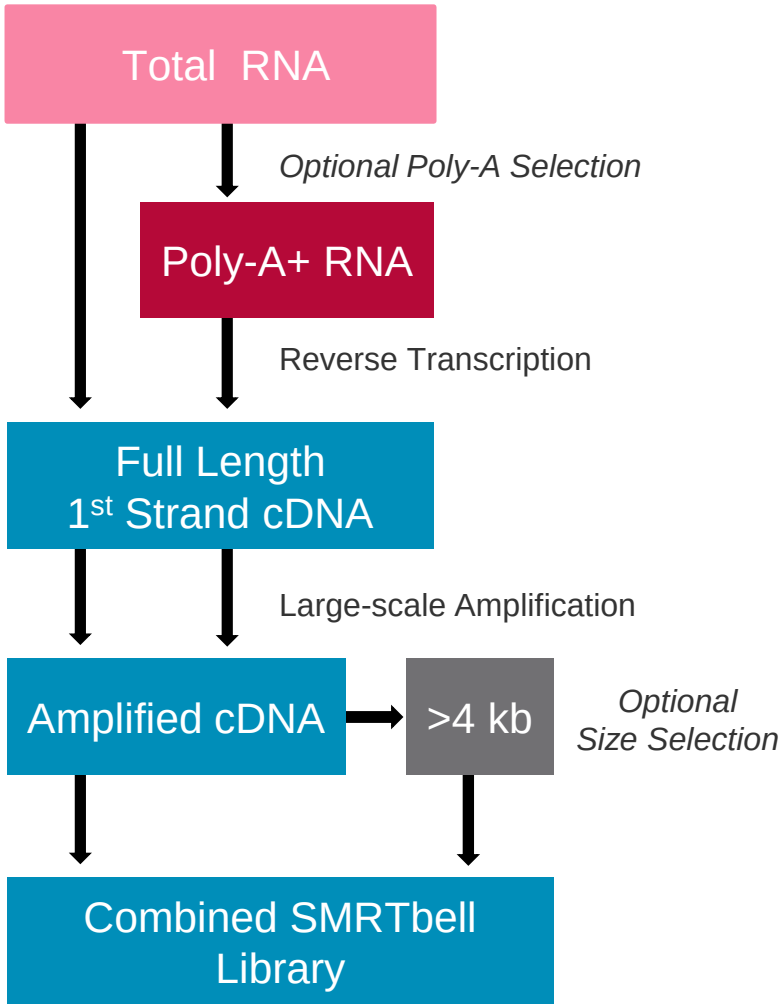
- Typically whole-transcriptome, non-quantitative
- Often included in *de novo* genome assembly projects
- Single tissue to several tissues
- Generates reference transcriptome for downstream RNA-seq studies

Gene-level Isoform Discovery

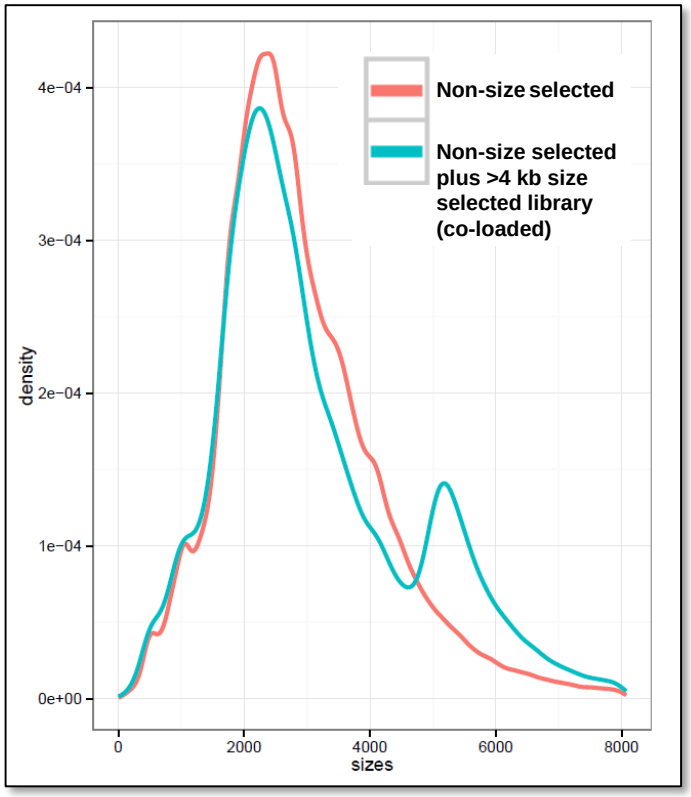
"Do alternative splicing or other transcription events play a role in a particular disease state?"

- Typically targeted, either cDNA amplicons or target capture
- Useful for detecting gene fusions, SNVs, allele-specific expression
- Cost-effectively multiplex many samples per single SMRT Cell
- Relative quantitation possible

SIMPLIFIED SEQUEL ISO-SEQ LIBRARY PREP



- Simplified library preparation
- Size selection optional



WHOLE-GENOME ANNOTATION: KEY PUBLICATIONS

Wang et al., **Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing**, *Nat Comm* (2016)

- First Iso-Seq application for whole genome annotation
- Multiplexed 6 different maize B73 tissues
- Obtained ~111 k high-quality transcripts
- Vastly improved existing annotation and incorporated to MaizeGDB v4

Wang et al., **A comparative transcriptional landscape of maize and sorghum obtained by single-molecule sequencing**, *Genome Research* (2018)

- Performed Iso-Seq method on maize and sorghum
- Comparative analysis of conserved and differentiated alternative splicing

WHOLE-GENOME ANNOTATION: KEY PUBLICATIONS

Kuo et al., **Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human.** *BMC Genomics* (2017)

- Whole transcriptome sequencing of chicken
- Used 5’ cap normalized Iso-Seq libraries
- Obtained ~60 k high-quality transcripts (~29 k genes)
- Identified >20 k potential lncRNAs

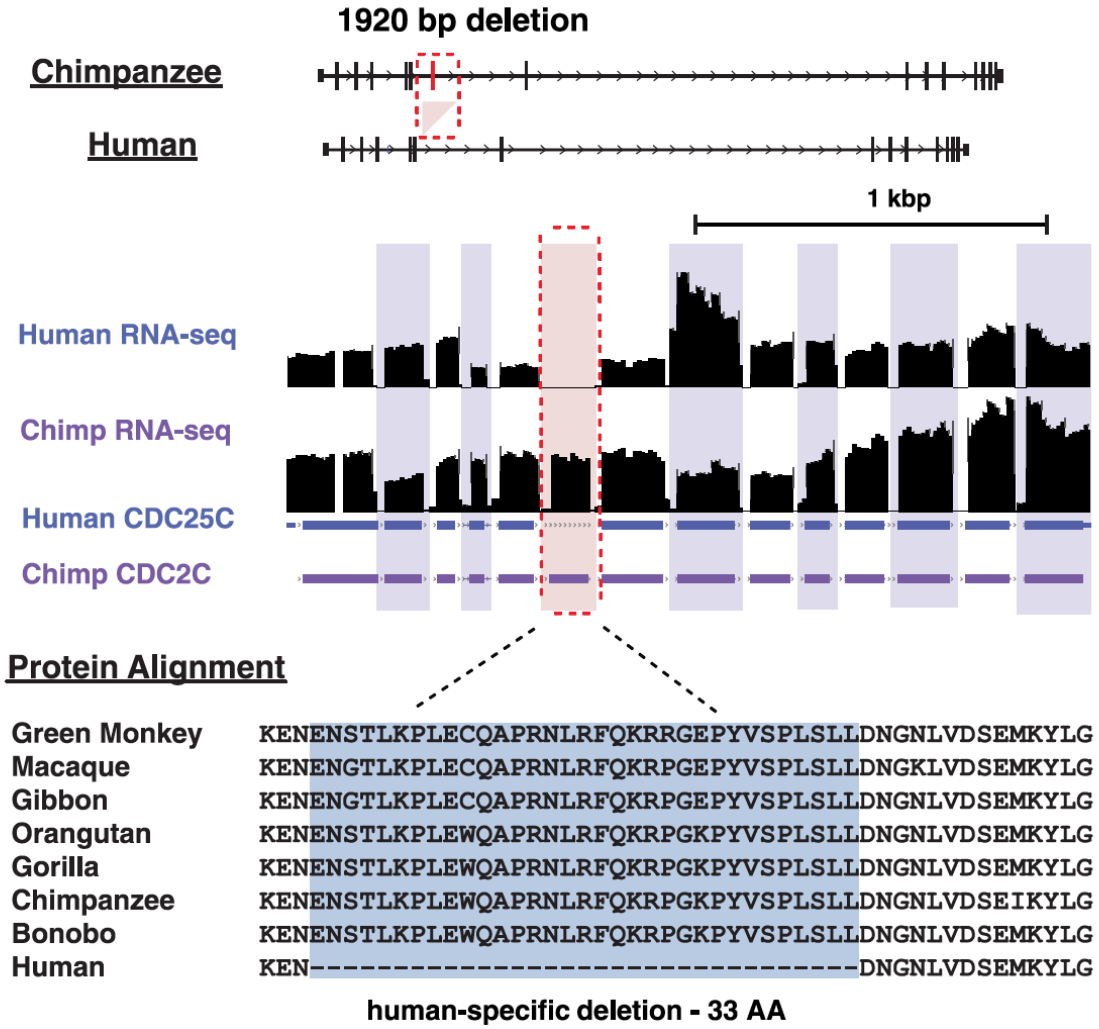
COMPARATIVE GENOME + TRANSCRIPTOME SEQUENCING

- Human, Chimp, and Orangutan
- *de novo* genome assembly using PacBio
- Iso-Seq + RNA-Seq for annotation

- Improved genome contiguity by 30- to 500-fold
- 83% of ape genome now in multi-species alignment
- Systematic SV discovery (~600 k in ape)
- Rare human-specific exonic deletion detected

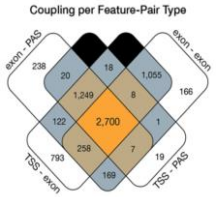
HUMAN SPECIFIC DELETIONS DETECTED BY CROSS-SPECIES ISO-SEQ COMPARISON

[Blog: Finding Human by sequencing our Ape relatives]



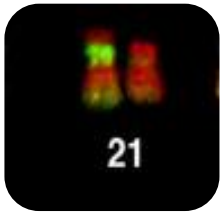
ISO-SEQ PUBLICATIONS: HUMAN GENES AND DISEASES

Treutlein et al., **Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing.** *Proc Natl Acad Sci* (2014)



Anvar et al., **Full-length mRNA sequencing uncovers a widespread coupling between transcription initiation and mRNA processing.** *Genome Biol.* (2018)

Kohli et al., **Androgen Receptor Variant AR-V9 Is Coexpressed with AR-V7 in Prostate Cancer Metastases and Predicts Abiraterone Resistance**, *Clinical Cancer Research* (2017)

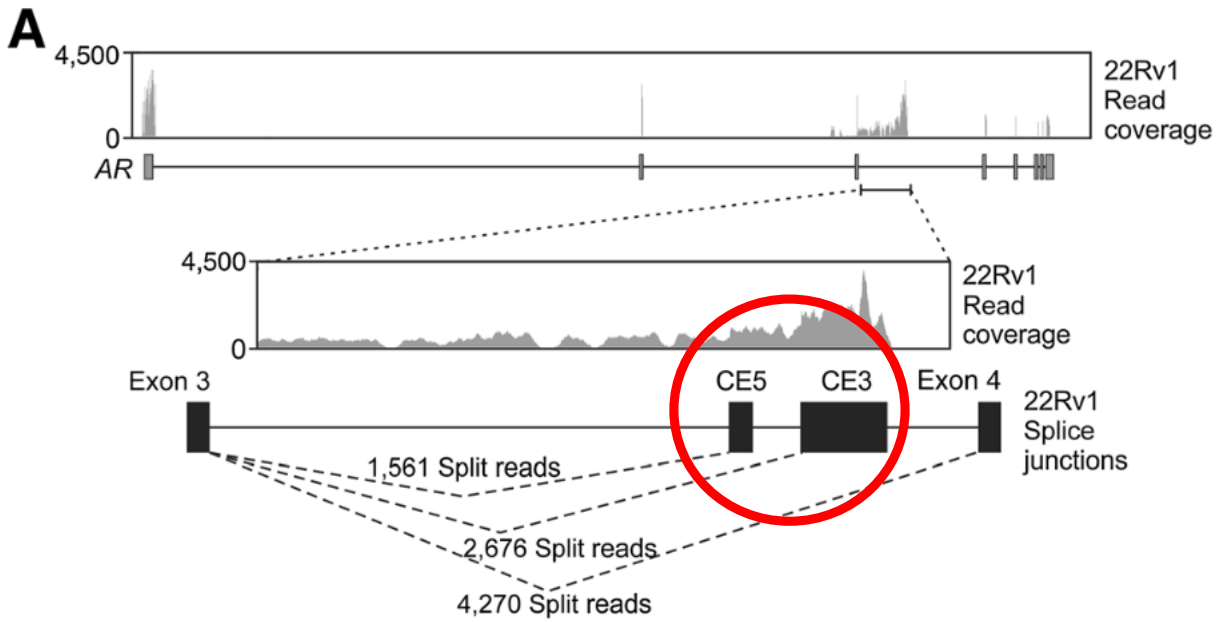


Deveson et al., **Universal Alternative Splicing of Noncoding Exons.** *Cell Systems* (2018)

Aneichyk et al., **Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly**. *Cell* (2018)

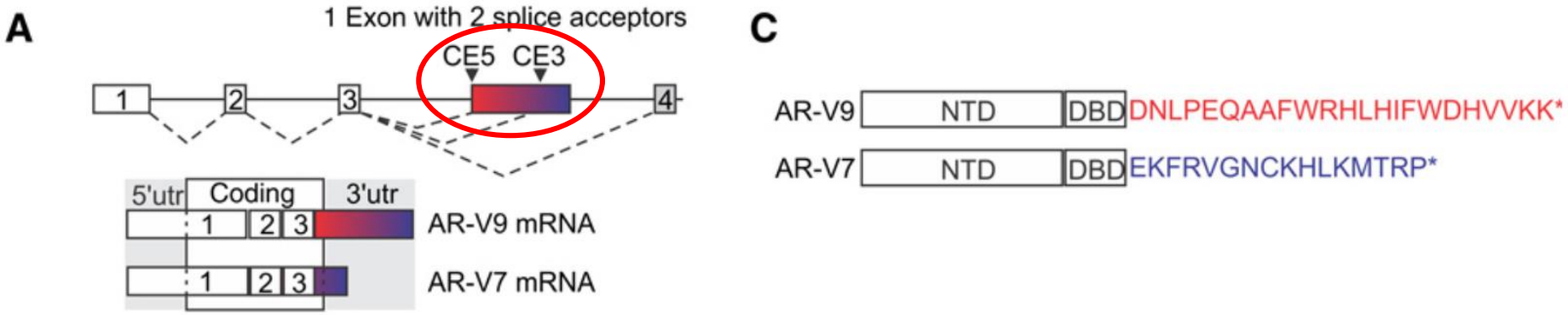
Kohli et al., **Androgen Receptor Variant AR-V9 Is Coexpressed with AR-V7 in Prostate Cancer Metastases and Predicts Abiraterone Resistance**, *Clinical Cancer Research* (2017)

- Sequenced only Andogren Receptor gene (AR) in prostate cancer
- AR-V7 is a known variant that prohibits successful therapy in castration-resistant prostate cancer



Kohli et al., **Androgen Receptor Variant AR-V9 Is Coexpressed with AR-V7 in Prostate Cancer Metastases and Predicts Abiraterone Resistance**, *Clinical Cancer Research* (2017)

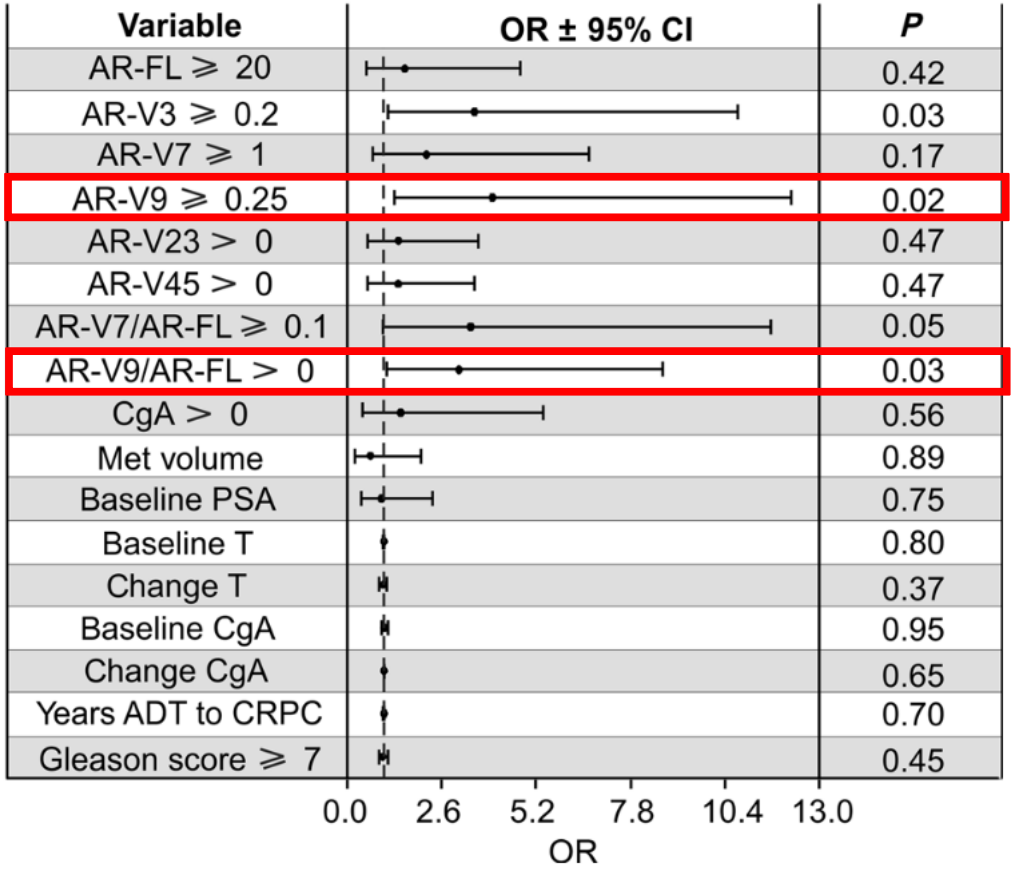
- Iso-Seq data identified AR-V9 often co-expressed with AR-V7
- Iso-Seq data re-annotated the cryptic exons CE3 and CE5 as a single 3' exon with different splice sites



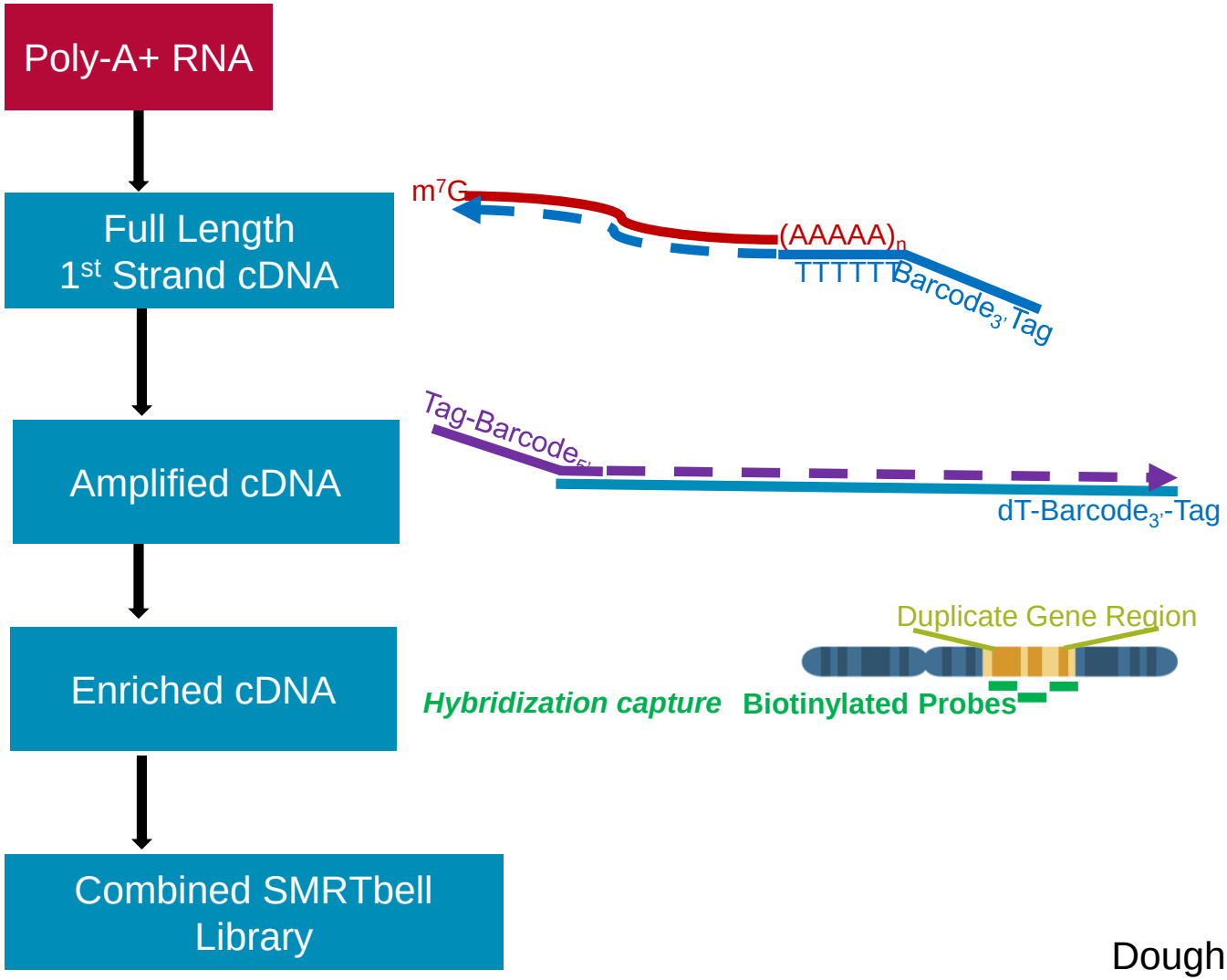
PROSTATE CANCER AWARENESS

Kohli et al., **Androgen Receptor Variant AR-V9 Is Coexpressed with AR-V7 in Prostate Cancer Metastases and Predicts Abiraterone Resistance**, *Clinical Cancer Research* (2017)

- AR-V9 expression predictive of therapy resistance



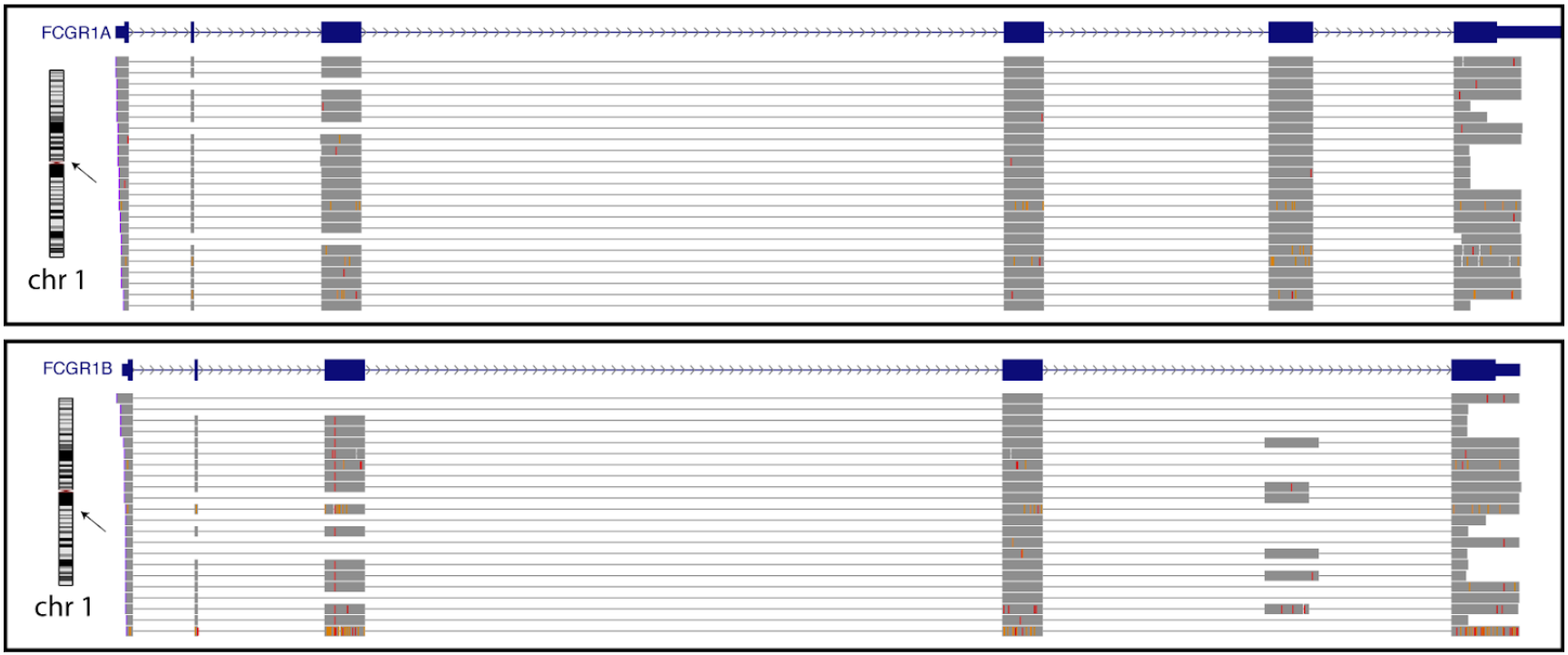
TARGETED ENRICHMENT OF SEGMENTAL DUPLICATED GENES



Dougherty et al. (accepted)

ISO-SEQ ANALYSIS CAPTURES SEGMENTAL DUPLICATED GENES

- FCGR1A and FCGR1B are > 99% similar



Dougherty et al. (accepted)

SINGLE-CELL APPLICATION

G&T-seq: parallel sequencing of single-cell genomes and transcriptomes

Iain C Macaulay[1], Wilfried Haerty[2,10], Parveen Kumar[3,10], Yang I Li[2,9], Tim Xiaoming Hu[2], Mabel J Teng[4], Mubeen Goolam[5], Nathalie Saurat[6], Paul Coupland[7], Lesley M Shirley[7], Miriam Smith[7], Niels Van der Aa[3], Ruby Banerjee[8], Peter D Ellis[7], Michael A Quail[7], Harold P Swerdlow[7,9], Magdalena Zernicka-Goetz[5], Frederick J Livesey[6], Chris P Ponting[1,2,11] & Thierry Voet[1,3,11]

The simultaneous sequencing of a single cell's genome and transcriptome offers a powerful means to dissect genetic variation and its effect on gene expression. Here we describe G&T-seq, a method for separating and sequencing genomic DNA and full-length mRNA from single cells. By applying G&T-seq to over 220 single cells from mice and humans, we discovered cellular properties that could not be inferred from DNA or RNA sequencing alone.

Karlsson and Linnarsson *BMC Genomics* (2017) 18:126
DOI 10.1186/s12864-017-3528-6

BMC Genomics

RESEARCH ARTICLE Open Access

Single-cell mRNA isoform diversity in the mouse brain

Kasper Karlsson[1] and Sten Linnarsson[2*]

Abstract

Background: Alternative mRNA isoform usage is an important source of protein diversity in mammalian cells. This phenomenon has been extensively studied in bulk tissues, however, it remains unclear how this diversity is reflected in single cells.

Results: Here we use long-read sequencing technology combined with unique molecular identifiers (UMIs) to reveal patterns of alternative full-length isoform expression in single cells from the mouse brain. We found a surprising amount of isoform diversity, even after applying a conservative definition of what constitutes an isoform. Genes tend to have one or a few isoforms highly expressed and a larger number of isoforms expressed at a low level. However, for many genes, nearly every sequenced mRNA molecule was unique, and many events affected coding regions suggesting previously unknown protein diversity in single cells. Exon junctions in coding regions were less prone to splicing errors than those in non-coding regions, indicating purifying selection on splice donor and acceptor efficiency.

Conclusions: Our findings indicate that mRNA isoform diversity is an important source of biological variability also in single cells.

CSH Cold Spring Harbor Laboratory

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

HOME | ABO

Search

New Results

Single-cell isoform RNA sequencing (ScISOr-Seq) across thousands of cells reveals isoforms of cerebellar cell types.

Ishaan Gupta, Paul G Collier, Bettina Haase, Ahmed Mahfouz, Anoushka Joglekar, Taylor Floyd, Frank Koopmans, Ben Barres, August B Smit, Steven Sloan, Wenjie Luo, Olivier Fedrigo, M Elizabeth Ross, Hagen U Tilgner

doi: https://doi.org/10.1101/364950

This article is a preprint and has not been peer-reviewed [what does this mean?].

Abstract Info/History Metrics Preview PDF

ADDITIONAL REFERENCES

Study Target	Approach	Publications
Single gene	cDNA amplicons Targeted enrichment	Tseng et al. (FMR1) Kohli et al. (AR) Aneichyk et al. (XDP)
10-200 genes	Targeted enrichment	Goldfeder (AGBT2018) Deveson et al. (chr21)
lncRNA	Normalization Targeted enrichment	Kuo et al. (chicken) Lagarde (GENCODE)
Differential expression	Combine with RNA-seq	Chen et al. (garlic)
Whole Transcriptome	Standard cDNA library	Anvar et al. (MCF-7)
Single Cell Sequencing	Combine with UMIs	Macaulay (G&T-Seq) Karlsson (mouse brain) Tilgner (ScISOr-Seq)

SEQUEL SYSTEM ISO-SEQ EXPERIMENT SIZE

SMRT Cells (per sample)	Experimental Goals
<1	Targeted, gene-specific isoform characterization
1	General survey of full-length isoforms in a transcriptome (moderate to high expression levels) with or without size selection
1-2	A comprehensive survey of full-length isoforms in the transcriptome (per sample)
2+	Deep sequencing for comprehensive isoform discovery and identification of low abundance transcripts (per sample)

Sequel Performance (5.1):

- Up to 20 Gb per SMRT Cell
- 20-hour movie time
- 250 kb - 350 kb full-length non-chimeric (FLNC) reads

Analysis:

- IsoSeq2 or Iso-Seq3 (beta) for whole-genome annotation and targeted experiments

Planned Improvements:

- IsoSeq3 in SMRT Link 6.0
- Up to 40 Gb per SMRT Cell (6.0)
- More high-quality long transcripts

www.pacb.com

Iso-Seq: How do we get there in the lab

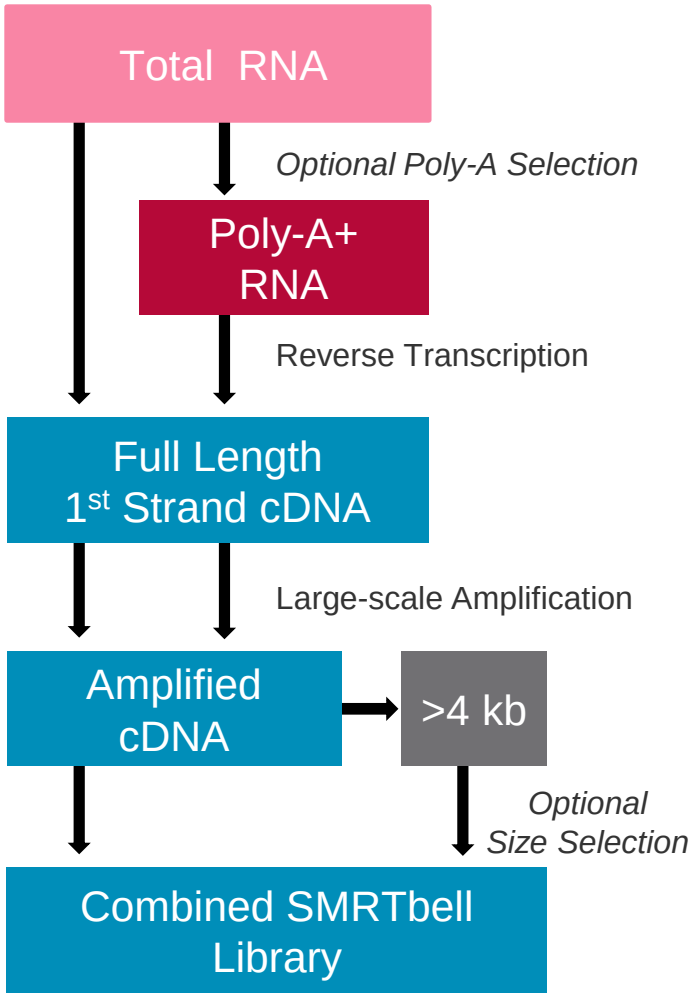
Pacific Biosciences User Group Meeting 2018

Michael Weiand

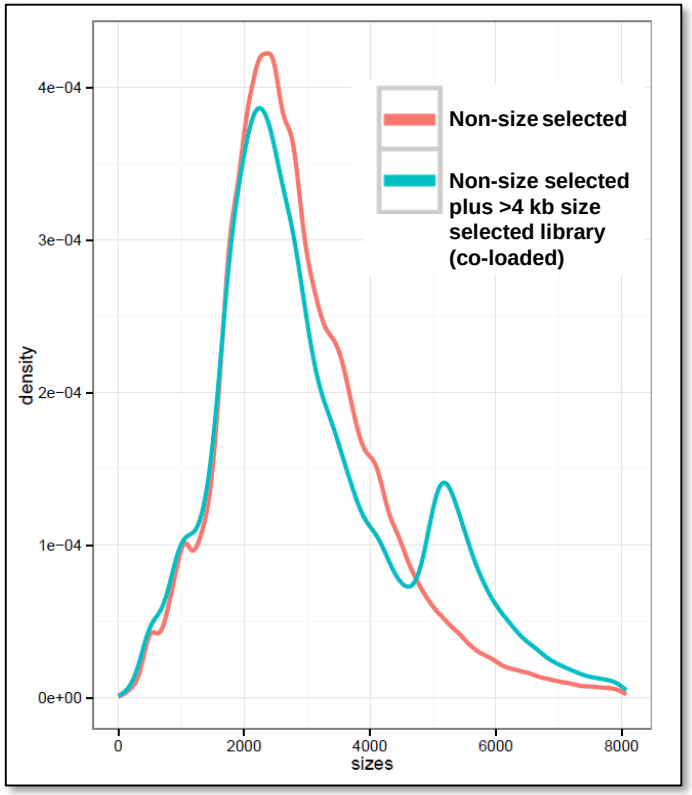
Breakout: **"The Why, What, and How of the Iso-Seq Method: Using full-length transcriptome sequencing to annotate genomes and solve diseases"**

- **Iso-Seq Sample Preparation Curent Updates**
  - **Customer Extraction Kits**
  - **Bacterial Iso-Seq for non-polyadenylated samples**
  - **10 hr vs 20 hr movies**
  - **Pre-Extension benefits**
  - **Upcoming Chemistry**

SEQUEL ISO-SEQ LIBRARY PREPARATION

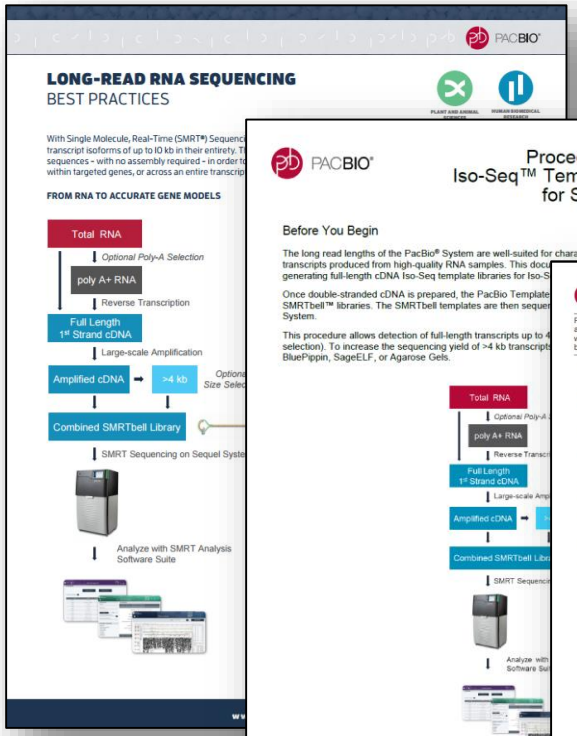


- Simplified library preparation
- Size selection optional

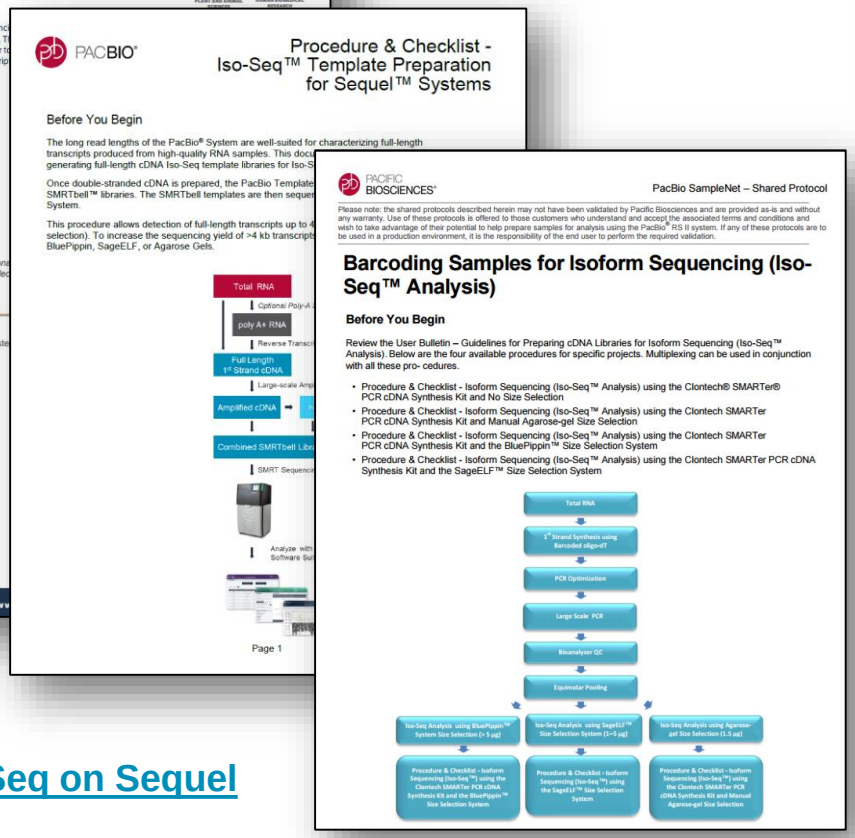


AVAILABLE TECHNICAL RESOURCES FOR ISO-SEQ ANALYSIS

Iso-Seq Best Practices

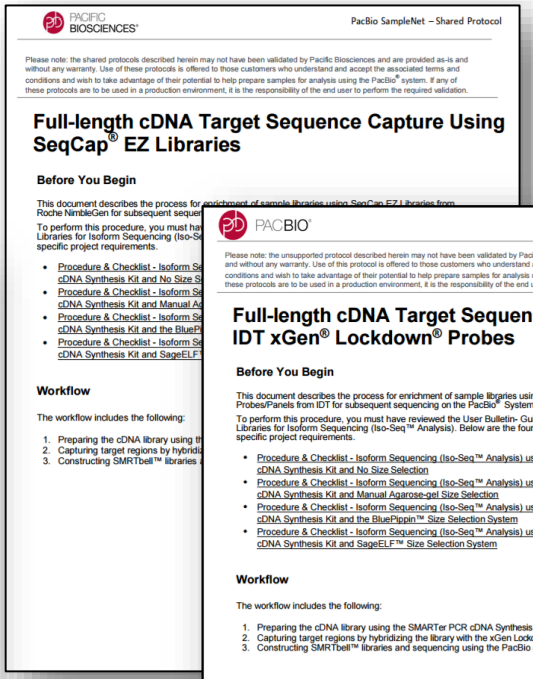


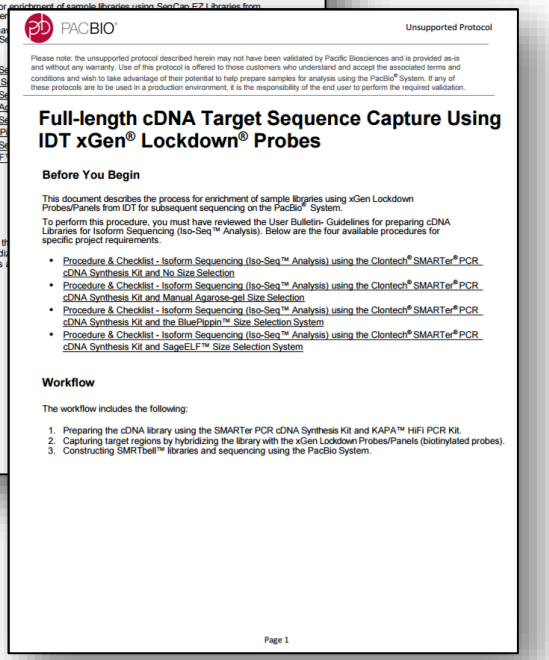
Barcoding Iso-Seq



Iso-Seq on Sequel

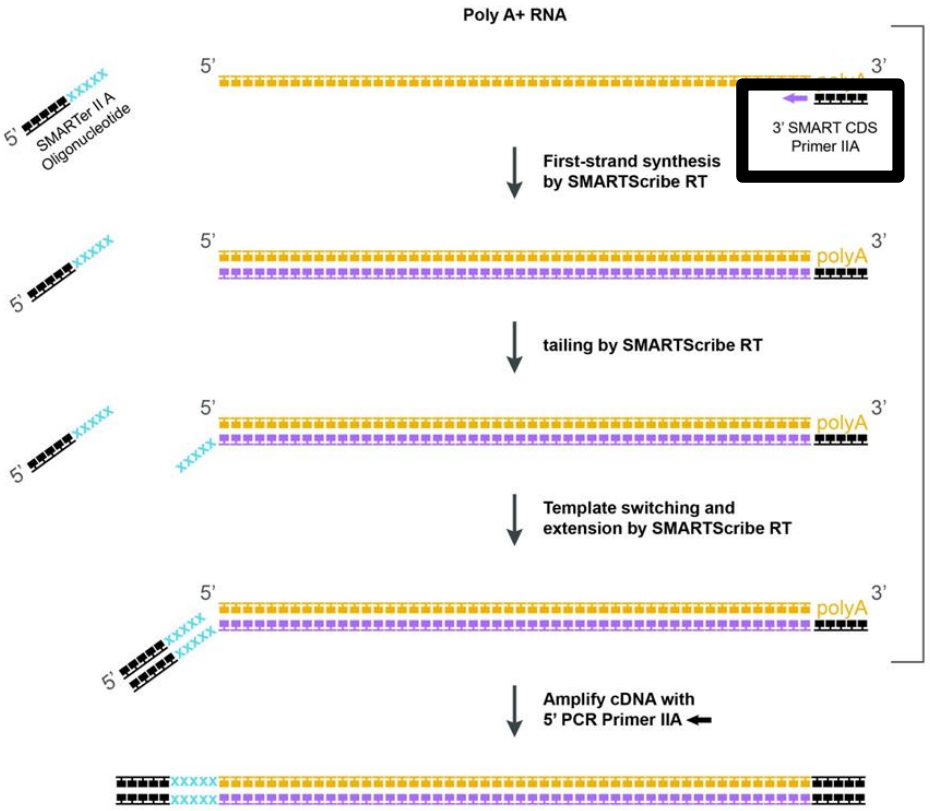
NimbleGen Targeted Iso-Seq





IDT Targeted Iso-Seq

BARCODING FOR ISO-SEQ ANALYSIS REQUIRES BARCODED OLIGO DT



Barcoded Oligo-dT

SMARTer_dT_BC1	AAGCAGTGGTATCAACGCAGAGTACtcagacgatgcgtcatTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
SMARTer_dT_BC2	AAGCAGTGGTATCAACGCAGAGTACctatacatgactctgcTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
SMARTer_dT_BC3	AAGCAGTGGTATCAACGCAGAGTACtactagagtagcactcTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
SMARTer_dT_BC4	AAGCAGTGGTATCAACGCAGAGTACtgtgtatcagtacatgTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
SMARTer_dT_BC5	AAGCAGTGGTATCAACGCAGAGTACgatctctactatatgcTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
SMARTer_dT_BC6	AAGCAGTGGTATCAACGCAGAGTACacagtctatactgctgTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN

- 24 validated barcodes for Sequel System
- Order oligos from any oligo synthesis providers
  - Barcode added between Poly-T and PCR priming site

BEFORE YOU START

Customer Used Extraction Kits

- Qiagen® RNeasy Plus Kits
  - www.qiagen.com/RNeasy
- Ambion® Poly(A) PuristTM MAG Kit
  - https://www.thermofisher.com/order/catalog/product/AM1922
- Sigma ® Spectrum Plant Total RNA Kit
- iNtRON Easy spin Total RNA
- RNALater which is a stabilizing storage solution. Also, any RNA prep solution where the nucleases are inactivated quickly (like TRIzol)

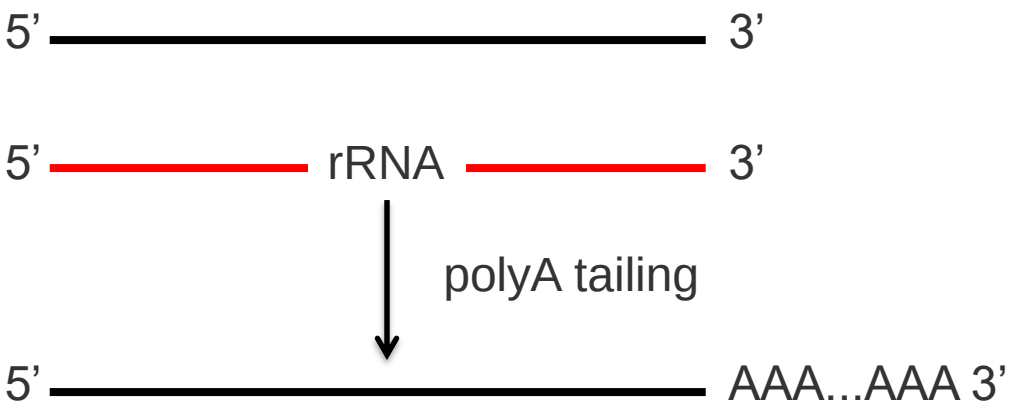
RNA Quality

- Has not been exposed to high temperatures (*e.g.*: > 65°C for 1 hour can cause a detectable decrease in sequence quality) or pH extremes (< 6 or > 9).
- Has an OD260/OD280 ratio between 2.0 and 2.2.
- Has an OD260/OD230 ratio between 1.8 and 2.1.
- Has a RIN number ≥ 8 (Recommended).
- Has not been exposed to intercalating fluorescent dyes or ultraviolet radiation. SYBR dyes are not RNA damaging, but do avoid ethidium bromide.
- Does not contain denaturants (e.g., guanidinium salts or phenol) or detergents (e.g., SDS or Triton-X100).
- Does not contain carryover contamination from the original organism/tissue (e.g., heme, humic acid, polyphenols, etc.).
- Note: RNA samples should only be shipped with dry ice.

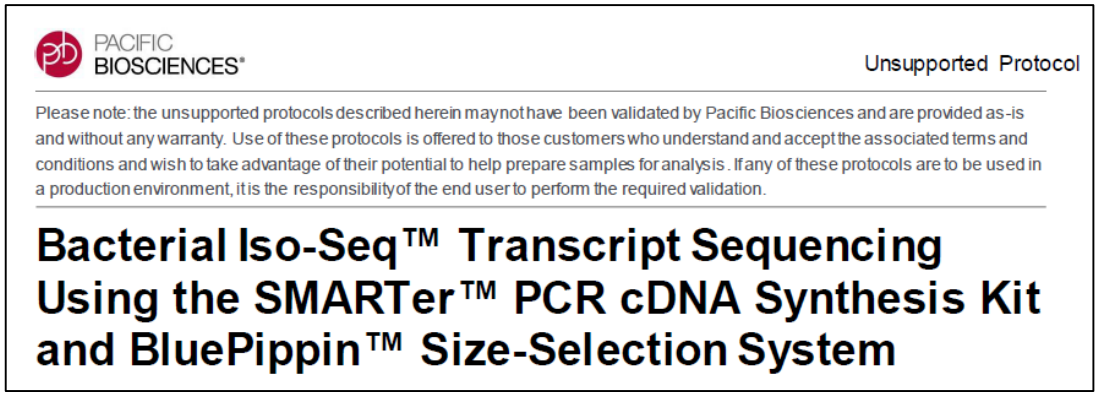
NO POLY-A??

Q: What if my sample doesn't have a Poly-A tail for RT priming with the Clontech kit?

A: Let's Enzymatically add it.



BACTERIAL ISO-SEQ FOR INCORPORATING POLY A



PACIFIC BIOSCIENCES®

Unsupported Protocol

Please note: the unsupported protocols described herein may not have been validated by Pacific Biosciences and are provided as-is and without any warranty. Use of these protocols is offered to those customers who understand and accept the associated terms and conditions and wish to take advantage of their potential to help prepare samples for analysis. If any of these protocols are to be used in a production environment, it is the responsibility of the end user to perform the required validation.

Bacterial Iso-Seq™ Transcript Sequencing Using the SMARTer™ PCR cDNA Synthesis Kit and BluePippin™ Size-Selection System

- Protocol available from PacBio
- http://www.pacb.com/wp-content/uploads/Unsupported-Protocol-Bacterial-Iso_Seq-Clontech-SMARTer-PCR-cDNA-Synthesis-Kit-BluePippin-Size-Selection.pdf

NEW ENGLAND BIOLABS PAPER

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-05997-6 OPEN

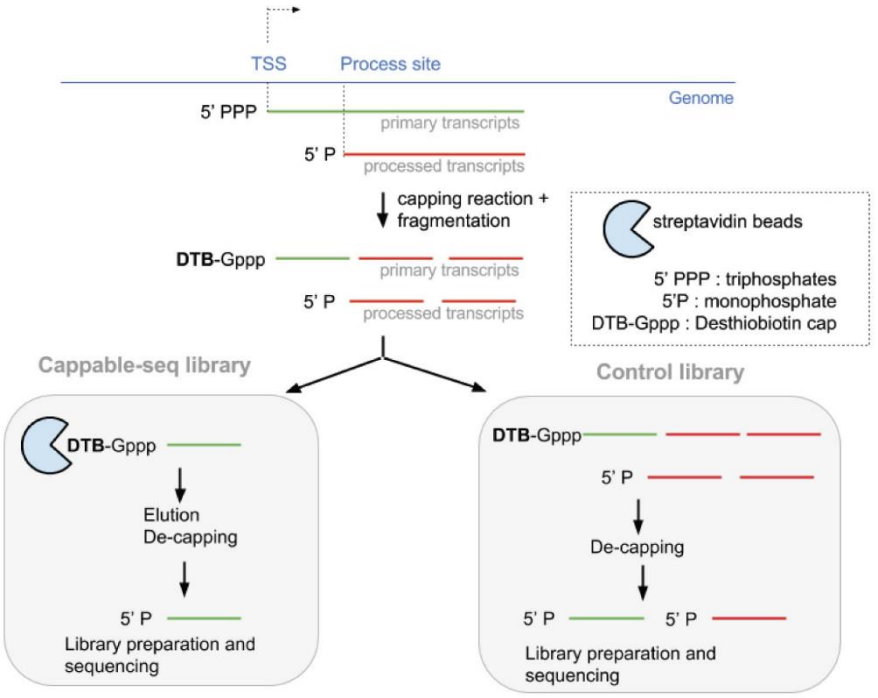
SMRT-Cappable-seq reveals complex operon variants in bacteria

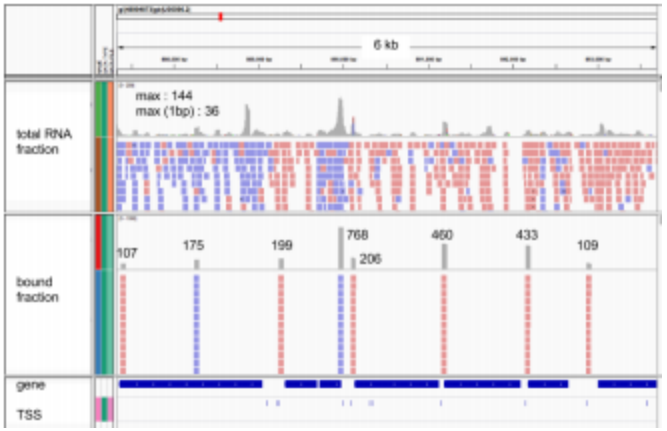
Bo Yan[1], Matthew Boitano[2], Tyson A. Clark[2] & Laurence Ettwiller[1]

- SMRT-Cappable-seq combines the isolation of full-length prokaryotic primary transcripts with long read sequencing technology
- “Pervasive read-through of previous experimentally validated transcription termination sites.”

SOLUTION: NEB CAPPABLE-SEQ

- NEB has developed an enrichment method to cap 5’ end of primary transcripts (non-processed – specific to 5’-PPP) with biotinylated GTP
- Most rRNA’s are not primary transcripts, they contain a processed 5’ end (5’-P)
- Protocol enables simultaneous rRNA depletion as well as enriching for primary transcripts with non-processed/non-degraded 5’ ends
  - Allowing for TSS and alternative TSS detection
- No bias observed between different organisms
- Current technologies impaired our ability to delineate transcript starts and ends that are typically several kb apart. Adopting methodology and incorporating long reads enable extended discovery

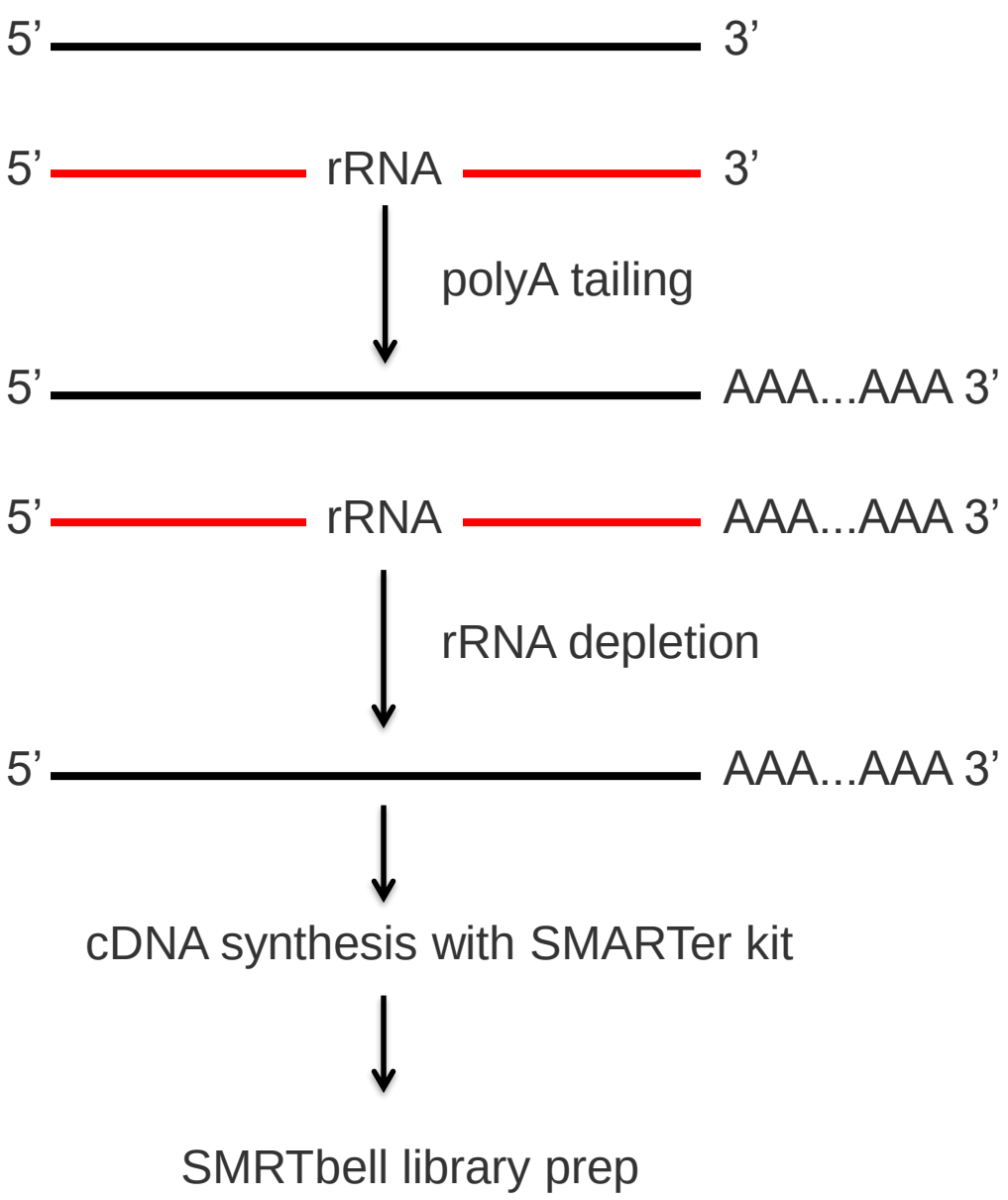




HIGH LEVEL PROTOCOL

Library Prep

- Begin with total RNA (of high quality)
  - Quality of data depends on input RNA quality
- Enzymatically add a polyA-tail
- Deplete rRNA using rRNA depletion kit:
  - Thermo (Invitrogen) RiboMinus Kit, Bacteria
  - NEB Cappable-Seq
  - Plug into standard Iso-Seq protocol



Iso-Seq Loading

LOADING FOR ISO-SEQ

Magbead loading on 2.1 chemistry

- Some customers have experienced variable and poor Magbead loading
- No identifiable root cause at this point

When to try Diffusion

- "**Observed** field magbead issues, still under investigation and limited resolutions
  - Focus shift to improve diffusion transcript lengths
    - Aim to get data consistently before and the right distribution of transcript lengths
    - Diffusion load to consistently get data, and use size selection strategies to get larger transcripts. E.g. you will consistently get data and we have a way to get the large transcripts.

ISO-SEQ LOADING – INTERNAL APPS LAB

SAMPLE	P1 (%)	Pol RL (mean,)	Insert Length (mean)	Pol Base Yield
90 pM, Magbead	4	41.5 kb	3.4 kb	1.5 Gb
10 pM, Diffusion	42	39.9 kb	1.9 kb	16.9 Gb

- 2.1 chemistry, both 4 hr pre-extension and 20 hr movie
- Side-by-side in the same run
- Good results from the Internal Control
- Better Longest Subread but low P1 by Magbeads

FOR DIFFUSION LOADING OF ISO-SEQ

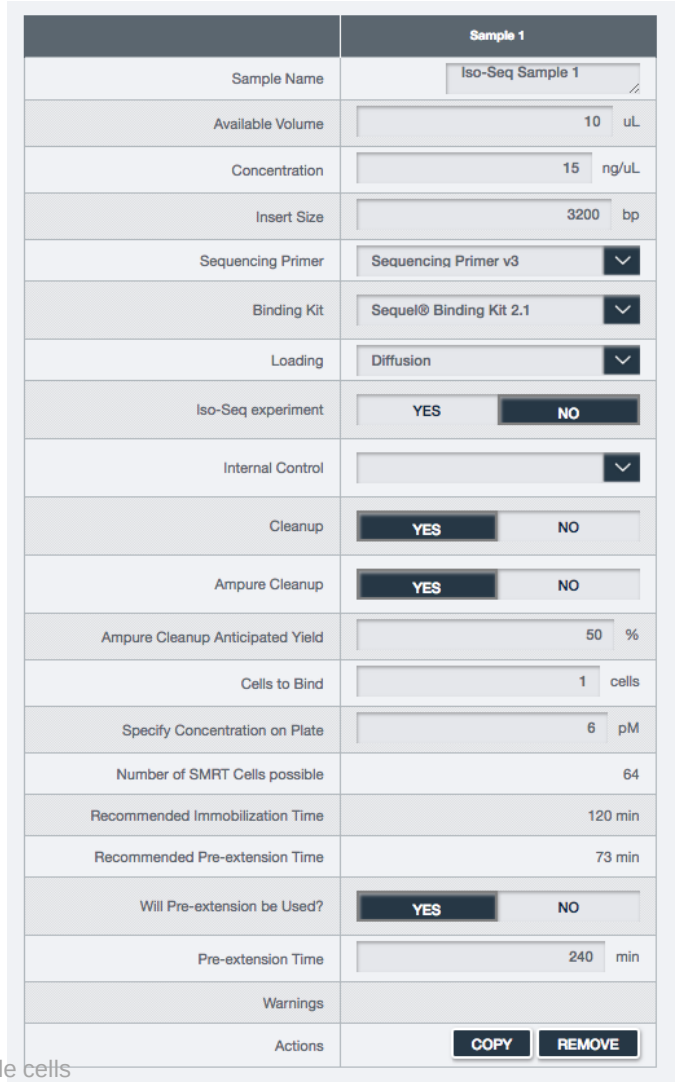
Sample Setup Guidance

Exceptions from standard Iso-Seq setup:

- Select **Diffusion** for Loading
- Select **NO** for Iso-Seq Experiment
- Select **YES** for Cleanup
- Select **YES** for AMPure Cleanup
- Enter 50% for AMPure Yield*

- Recommended on-plate concentration is 2-8 pmol, though higher amounts might be necessary for some samples.

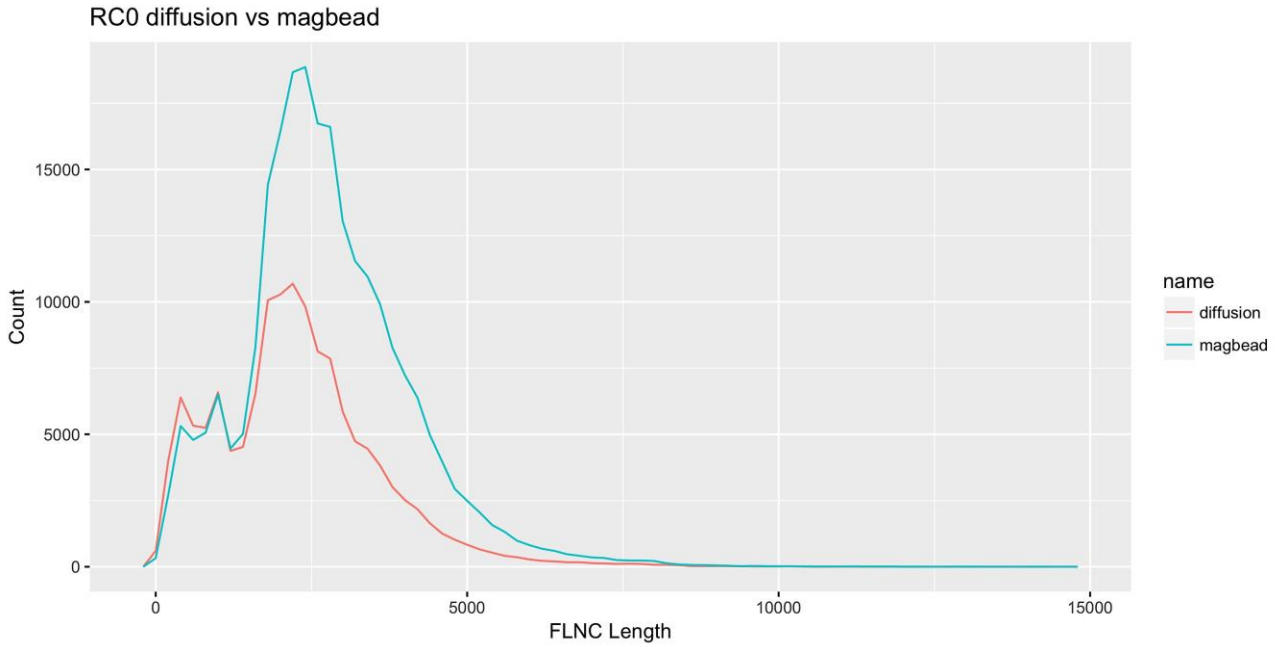
- Target P1 up to 70%, and P2<20%.

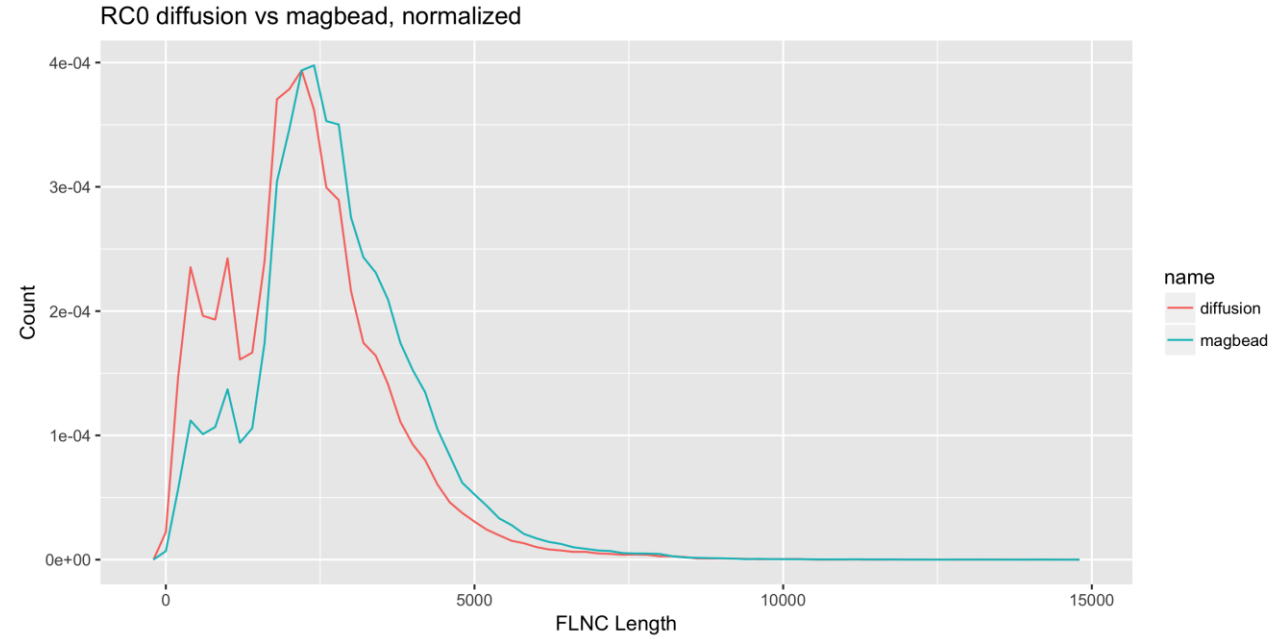


	Sample 1
Sample Name	Iso-Seq Sample 1
Available Volume	10 uL
Concentration	15 ng/uL
Insert Size	3200 bp
Sequencing Primer	Sequencing Primer v3
Binding Kit	Sequel® Binding Kit 2.1
Loading	Diffusion
Iso-Seq experiment	YES / NO
Internal Control	
Cleanup	YES / NO
Ampure Cleanup	YES / NO
Ampure Cleanup Anticipated Yield	50 %
Cells to Bind	1 cells
Specify Concentration on Plate	6 pM
Number of SMRT Cells possible	64
Recommended Immobilization Time	120 min
Recommended Pre-extension Time	73 min
Will Pre-extension be Used?	YES / NO
Pre-extension Time	240 min
Warnings	
Actions	COPY / REMOVE

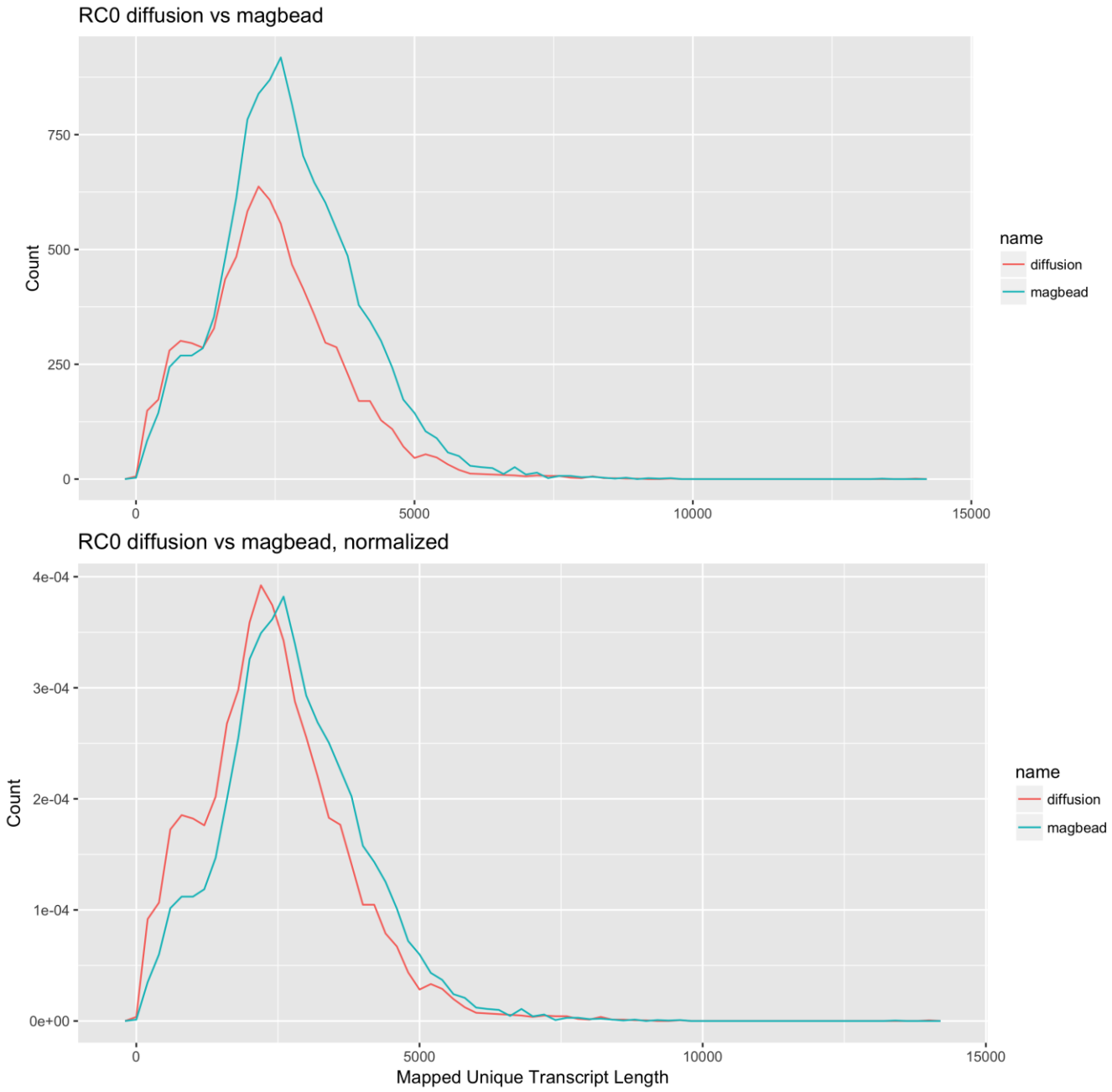
*Expect low AMPure recovery but will still have enough complex remaining to load multiple cells

LENGTH COMPARISON: FLNC READS





LENGTH COMPARISON: MAPPED UNIQUE TRANSCRIPT



Example I

BENEFITS OF PRE-EXTENSION FOR ISO-SEQ

Sequencing through the second adapter and back onto the initial strand:

- Polymerase activity is most stable during rolling circle replication mode
  - Start data collection **here**
- Pre-extension time value depends on the insert size

- *Pre-extension (PE) enables the start of movie acquisition to be **delayed** until the polymerase enters rolling circle replication and is in its **most stable** (processive) phase of sequencing*

- Pre-extension helps increase mean Polymerase Read Length metric by reducing the number of early-termination reads collected during primary analysis
- For Iso-Seq, recommended pre-extension time is **240 minutes**

PACBIO®

Diffusion Loading and Pre-Extension Time Recommendations for the Sequel® System Quick Reference Card

Mean Insert Size	Pre-Extension Time (minutes)
< 1 kb	N/A
1 kb	> 30
2 kb	45
4 kb	90
6 kb	135
10 kb	220
Iso-Seq	240
Size Selected Libraries	N/A

MODERATE GAINS IN GENE AND ISOFORM DETECTION AS MOVIE TIME AND PRE-EXTENSION ARE INCREASED

COLLECTION DESCRIPTION	# OF FULL LENGTH	# OF GENES	% INCREASE IN GENES	# OF ISOFORMS	% INCREASE IN ISOFORMS
10h movie, 2h pre	182,823	6645	BASE	11,095	BASE
20h movie, 2h pre	209,459	6955	5%	12,138	9%
20h movie, 4h pre	269,074	7933	19%	16,053	45%

- Full Length (FL) reads are determined by presence of 5’ / 3’ cDNA primer and polyA tail
- Number of genes/isoforms determined after running Iso-Seq3 analysis and mapping HQ isoform sequences to hg38, then categorizing it using SQANTI

What’s coming…

Sequel 6.0 Release Preview

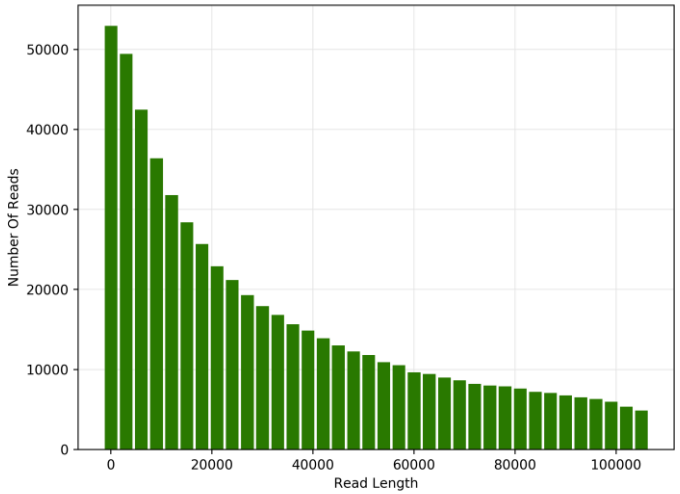
CHEMISTRY 3.0 FOR ISO-SEQ

	P1 (%)	Pol RL (mean)	Insert Length (mean)	Pol Base Yield
10 hr 4.3pM cell 1	62	35862	3491	22 Gb
10 hr 4.3pM cell 2	64	30621	3510	19 Gb
20 hr 4.3pM	69	46142	3989	31 Gb
20 hr 5pM	72	39465	4238	28 Gb

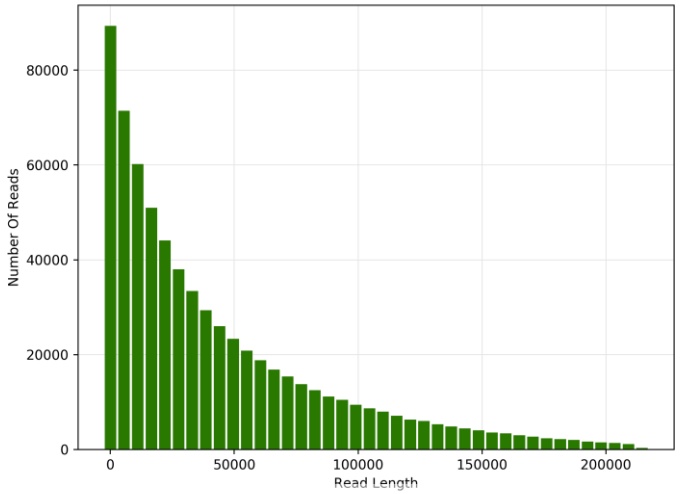
- All samples are RC0 (human universal reference)
- All samples were prepped using current Iso-Seq library protocol
- All samples were run with **diffusion loading** using 3.0 chemistry

POLYMERASE READ LENGTH: 10 HR VS 20 HR

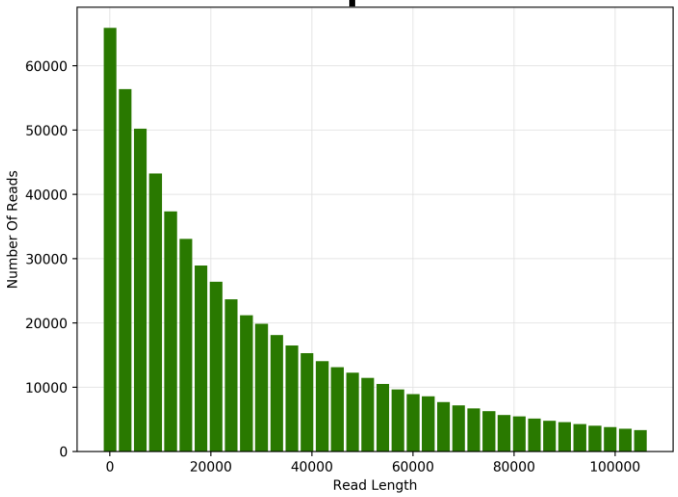
10 hr 4.3pM cell 1



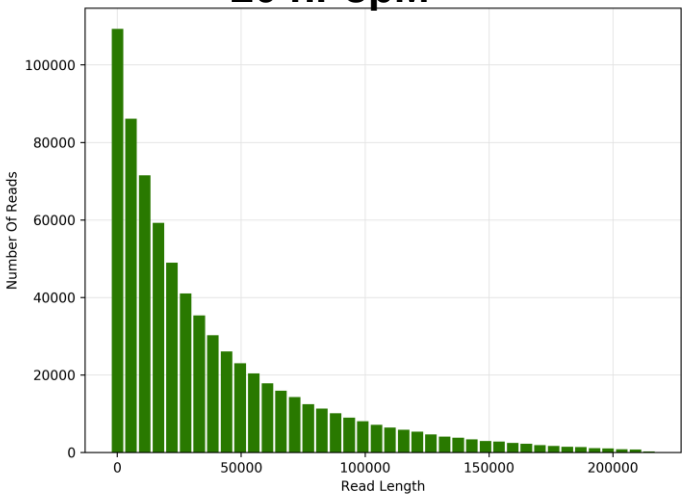
20 hr 4.3pM



10 hr 4.3pM cell 2



20 hr 5pM



FULL-LENGTH CCS READS: 10 HR VS 20 HR

	CCS	FLNC	FLNC%	FLNC RL (mean)
10 hr 4.3pM cell 1	499,933	376,056	75%	2.5 kb
10 hr 4.3pM cell 2	503,045	372,253	74%	2.5 kb
20 hr 4.3pM	572,406	430,257	75%	2.5 kb
20 hr 5pM	575,940	420,709	73%	2.5 kb

MAPPED TRANSCRIPTS: 10 HR VS 20 HR

	HQ Transcripts	Mapped Unique Genes*	Mapped Unique Transcripts*
10 hr 4.3pM cell 1	25,692	9269	19746
10 hr 4.3pM cell 2	25,208	9256	19448
20 hr 4.3pM	28,579	9795	21655
20 hr 5pM	27,791	9574	20997

* Unique genes and transcripts determined after mapping HQ transcripts to hg38 and compared with Gencode v27

SUMMARY

- Prepare full-length transcripts using the Clontech® SMARTer® PCR cDNA Synthesis Kit with as little as 1 ng of poly A+ RNA or 2 ng of total RNA
- Sequel System loading protocols reduce need for size selection for transcripts <4 kb
  - Optional size-selection protocols to enrich for transcripts >4 kb
- Compatible with standard target enrichment methods, such as NimbleGen SeqCap EZ or IDT xGen Lockdown Probes
- Multiplex samples to reduce sequencing needs
- Data analysis protocols and tools available through SMRT Analysis and Bioconda to generate high-quality, full-length transcript sequences with no assembly required

www.pacb.com

Iso-Seq: Introduction and Applications

NA UGM 2018- Iso-Seq Breakout Session

Google Group:

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

GitHub Repository and Tutorials:

github.com/PacificBiosciences/IsoSeq_SA3nUP/

(**http://tinyurl.com/PBisoseq**)

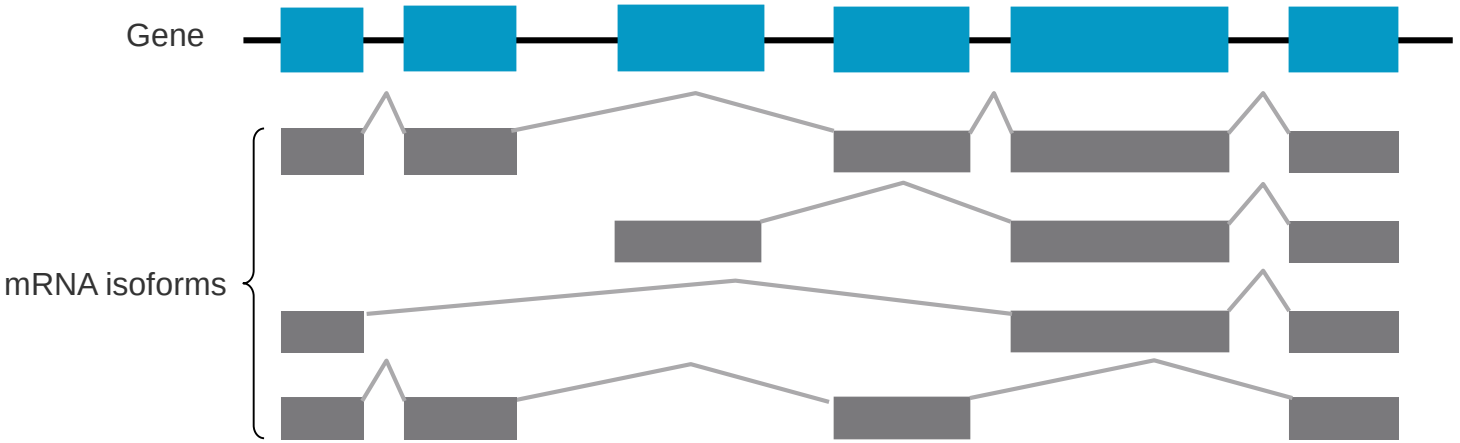
ISO-SEQ OVERVIEW

- Iso-Seq (“Isoform Sequencing”) is the umbrella term of transcriptome sequencing and downstream analysis using PacBio

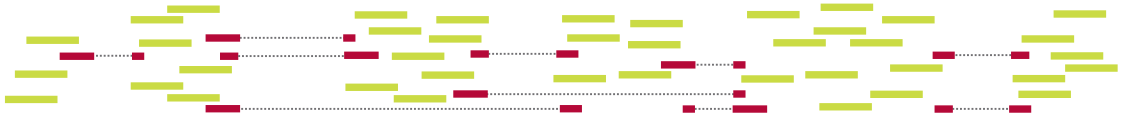
- Applications include:
  - whole genome annotation
  - isoform discovery
  - fusion gene detection
  - creating *de novo* reference transcripts for RNA-seq quantification

- In this session, we will:
  - 1) Review sequencing coverage recommendations
  - 2) Review the general Iso-Seq informatics workflow
  - 3) Discuss downstream tools for biological interpretation
  - 4) Preview the Iso-Seq3 workflow, including benchmarking and performance

DETERMINATION OF TRANSCRIPT ISOFORMS



Short-read technologies:



Insufficient Connectivity
Splice Isoform Uncertainty

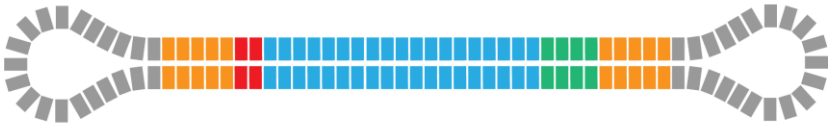
Reads spanning splice junctions

PacBio Iso-Seq solution:

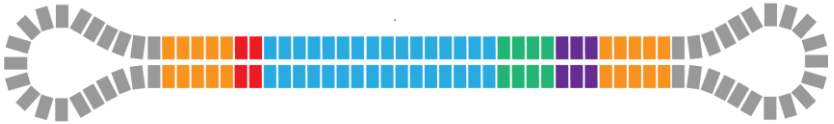
Full-length cDNA Sequence Reads
Spliced Isoform Certainty – No Assembly Required

ISO-SEQ SUPPORTS VARIOUS EXPERIMENTAL SETUPS

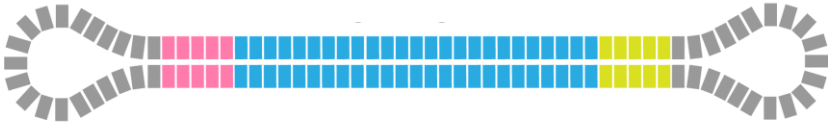
Whole transcriptome

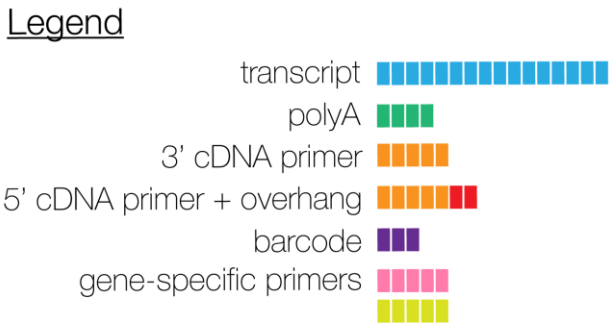


Whole transcriptome, barcoded



Targeted genes





PACBIO®

Considerations for Sequencing Coverage

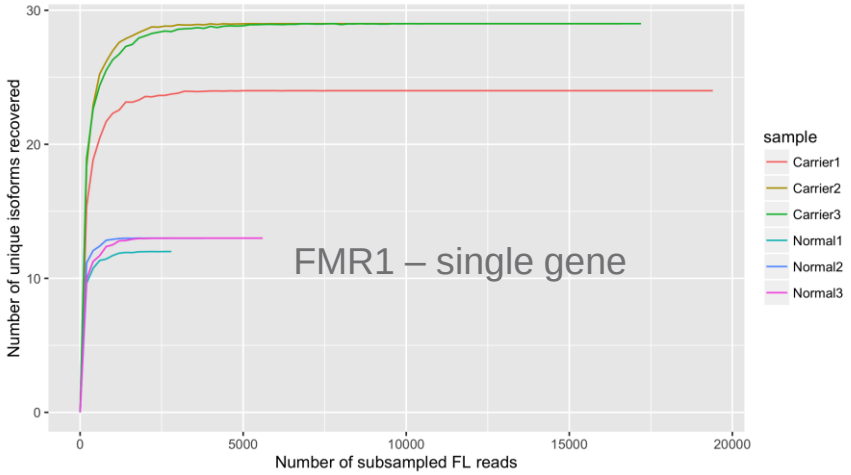
ISO-SEQ AT SEQUEL-SCALE

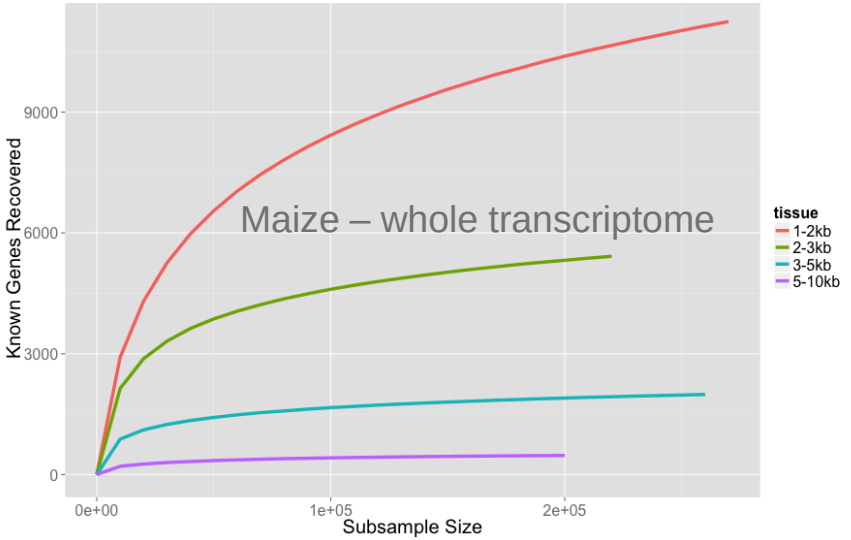
Targeted Genes:

- < 1 Sequel Cell
- Multiplexing Recommended

Whole Transcriptome:

- 2 – 4 Sequel Cell
- Multiplexing Recommended





Tseng et al., **Altered expression of the FMR1 splicing variants landscape in premutation carriers**, to appear in BBA – Gene Regulatory Mechanisms (2017)

Wang et al., **Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing**, *Nat Comm* (2016)

GENOME ANNOTATION AT SEQUEL SCALE

	NUMBER OF FL READS	NUMBER OF GENES	NUMBER OF ISOFORMS	Would be:
Maize	1,553,692	26,946	111,151	~6 Sequel Cell
Chicken	653,441	29,013	64,277	~3 Sequel Cell
Rabbit	466,034	14,474	36,186	~2 Sequel Cell
R. necatrix	330,373	> 5000	10,616	~2 Sequel Cell
Zebra Finch	405,736	7,228	17,437	Actual ~2 Sequel Cell

Wang et al., **Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing**, *Nat Comm* (2016)

Kuo et al., **Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human**, *BMC Genomics* (2017)

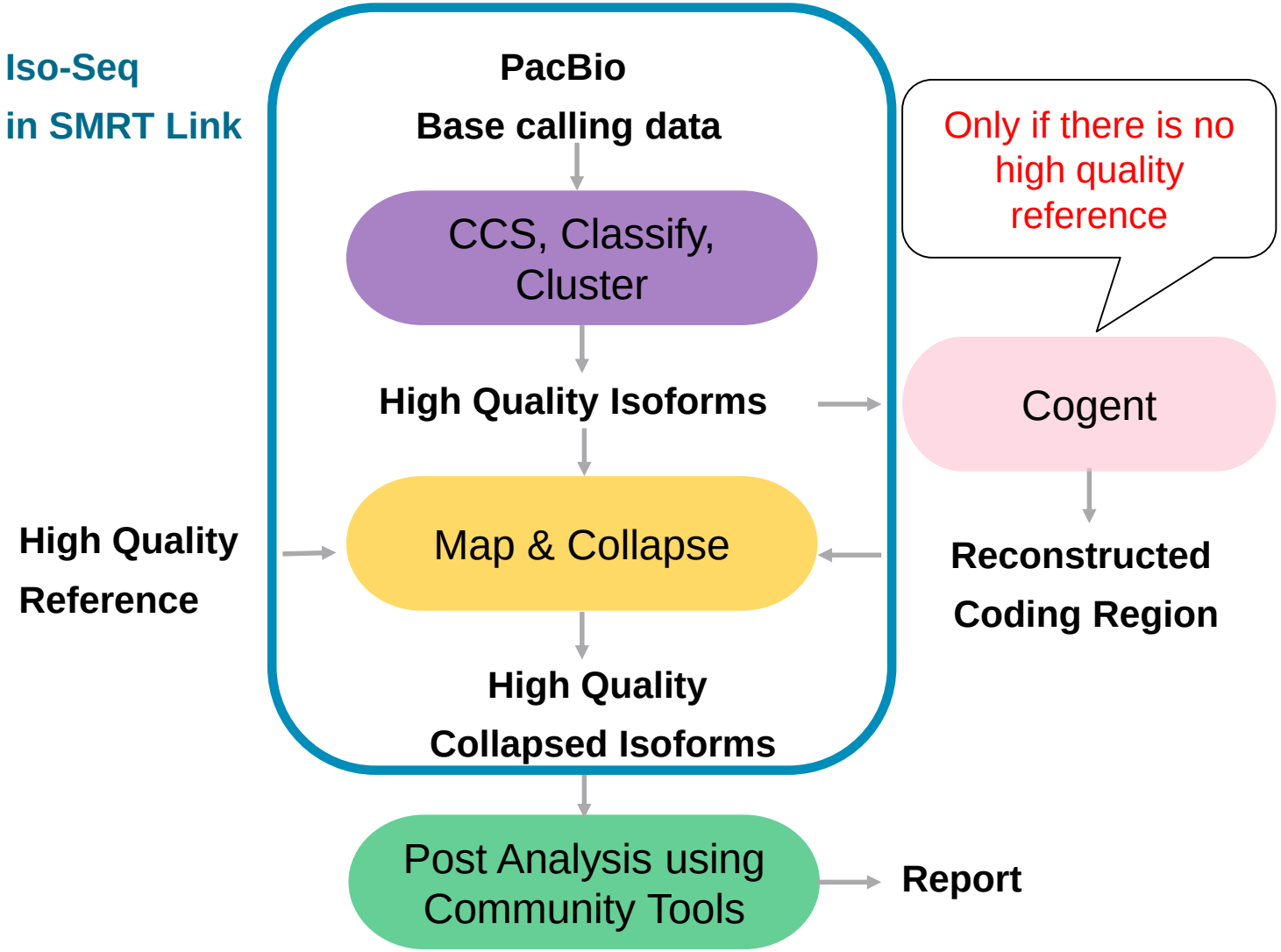
Chen et al., **A transcriptome atlas of rabbit revealed by PacBio single-molecule long-read sequencing**, *Sci Rep* (2017)

Kim et al.,**Characterization of the Rosellinia necatrix Transcriptome and Genes Related to Pathogenesis by Single-Molecule mRNA Sequencing,** *Plant Patho J* (2017)

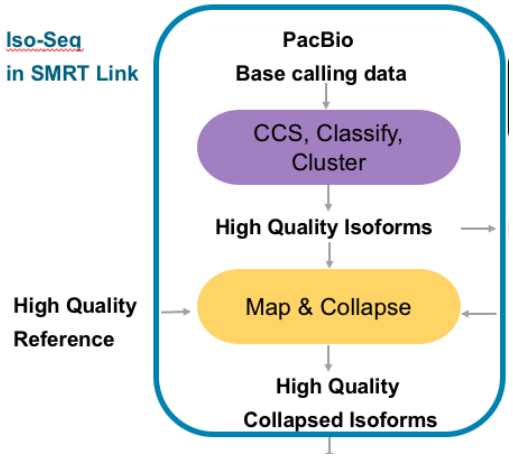
PACBIO®

Bioinformatics Deep Dive

ISO-SEQ ANALYSIS WORKFLOW



ISO-SEQ: FULL-LENGTH TRANSCRIPT SEQUENCING



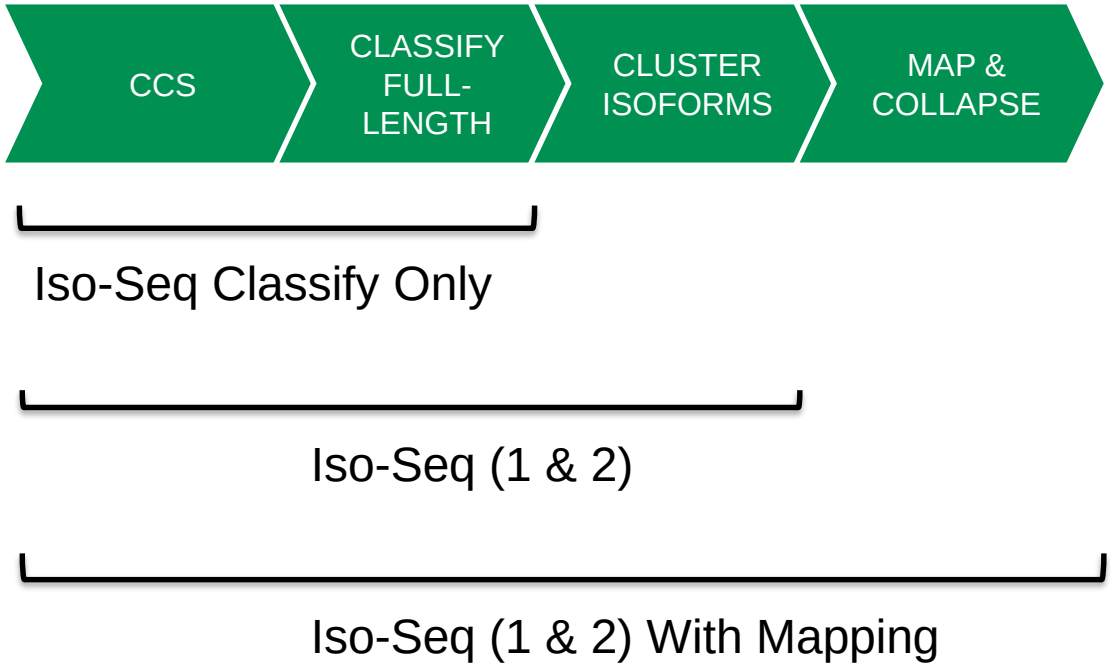
GENERATE CCS

CLASSIFY FULL-LENGTH

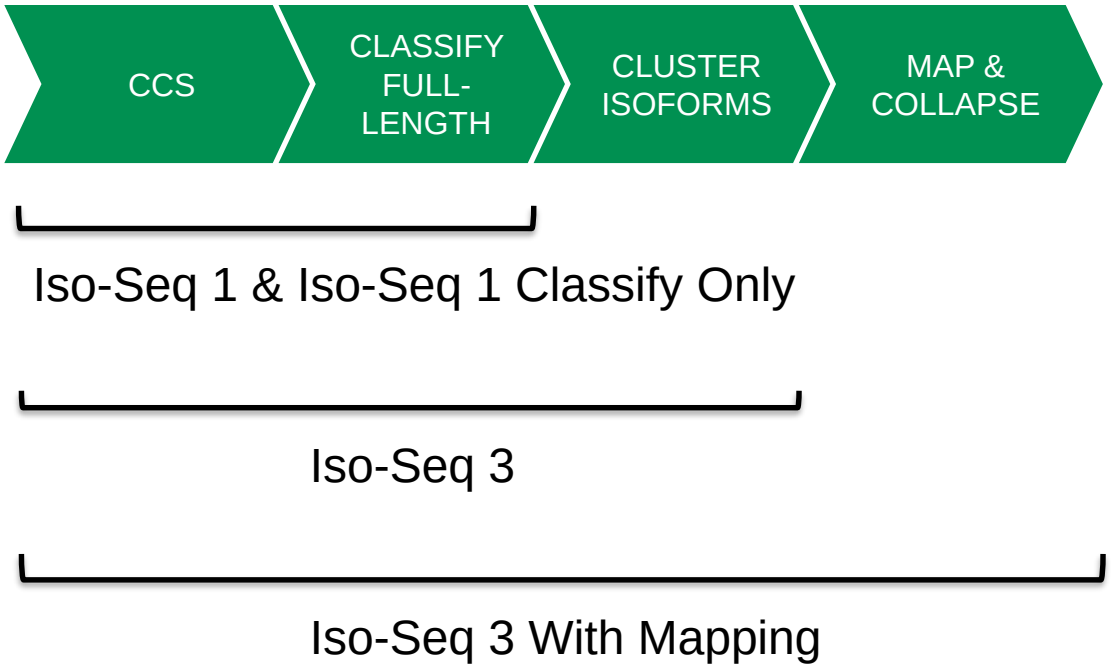
CLUSTER ISOFORMS

MAP & COLLAPSE

CURRENTLY AVAILABLE PIPELINES: SMRT LINK 5.1



FUTURE: ISO-SEQ WORKFLOWS IN SMRT LINK V6.0



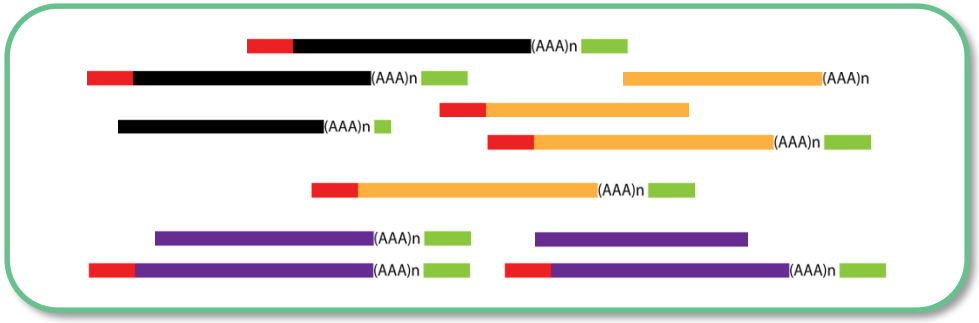
Iso-Seq 1 and Iso-Seq 1 With Mapping will be obsoleted in the future

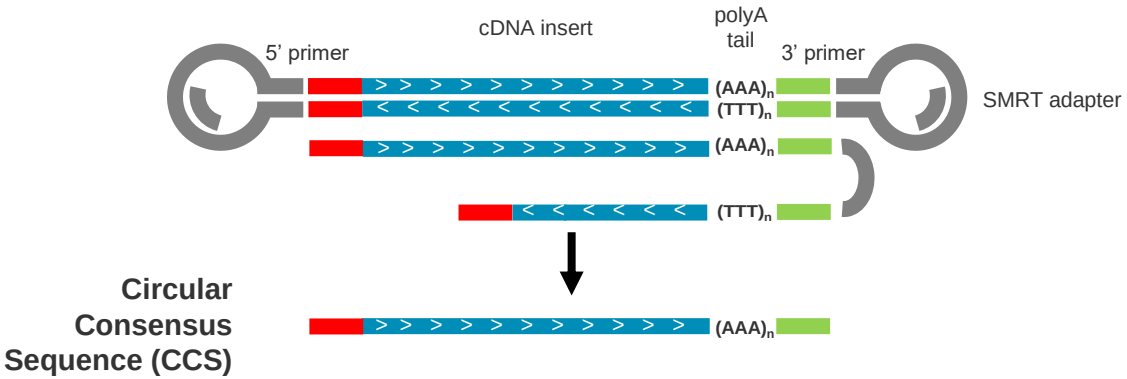
Iso-Seq 2 and Iso-Seq 2 with Mapping are already removed in SMRT Link v6.0

More information on SMRT Link 6.0 featured in "Evolving SMRT Applications" Breakout session

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

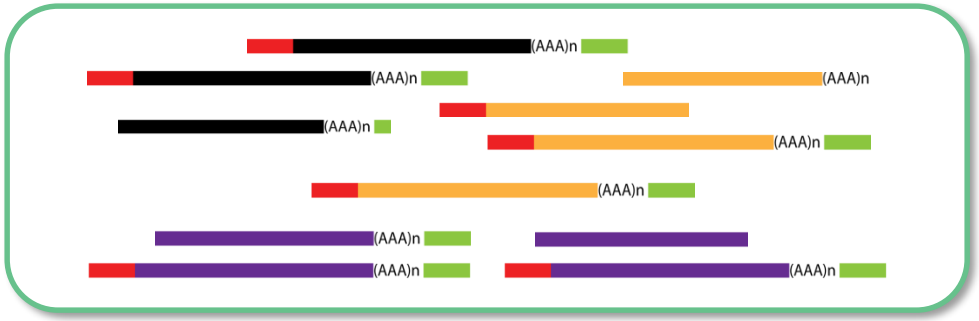
CCS





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

CCS



nFL reads

Iso-Seq 1 & 2

FL reads

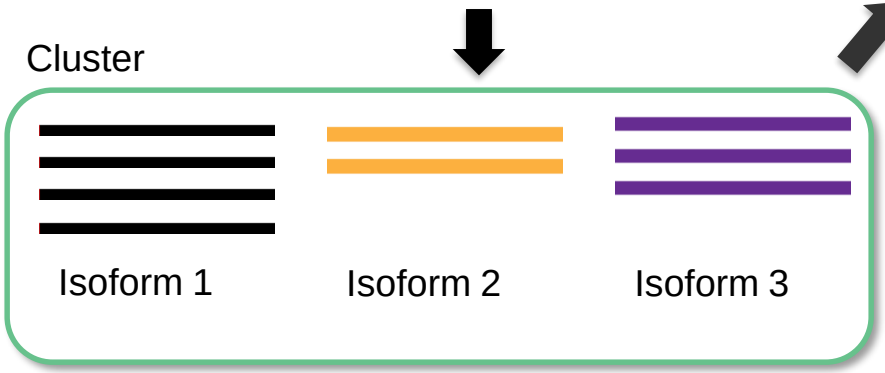
Full-length:

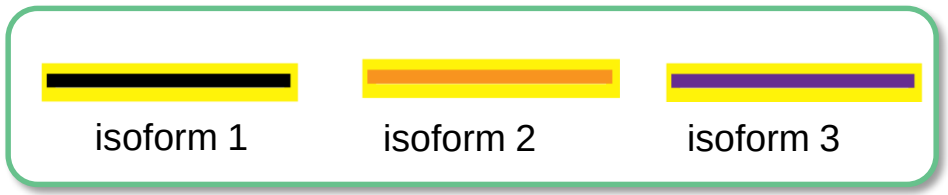
- Has 5’ cDNA primer
- Has 3’ cDNA primer
- Has polyA tail (>20 bp)

Support custom library prep and FL

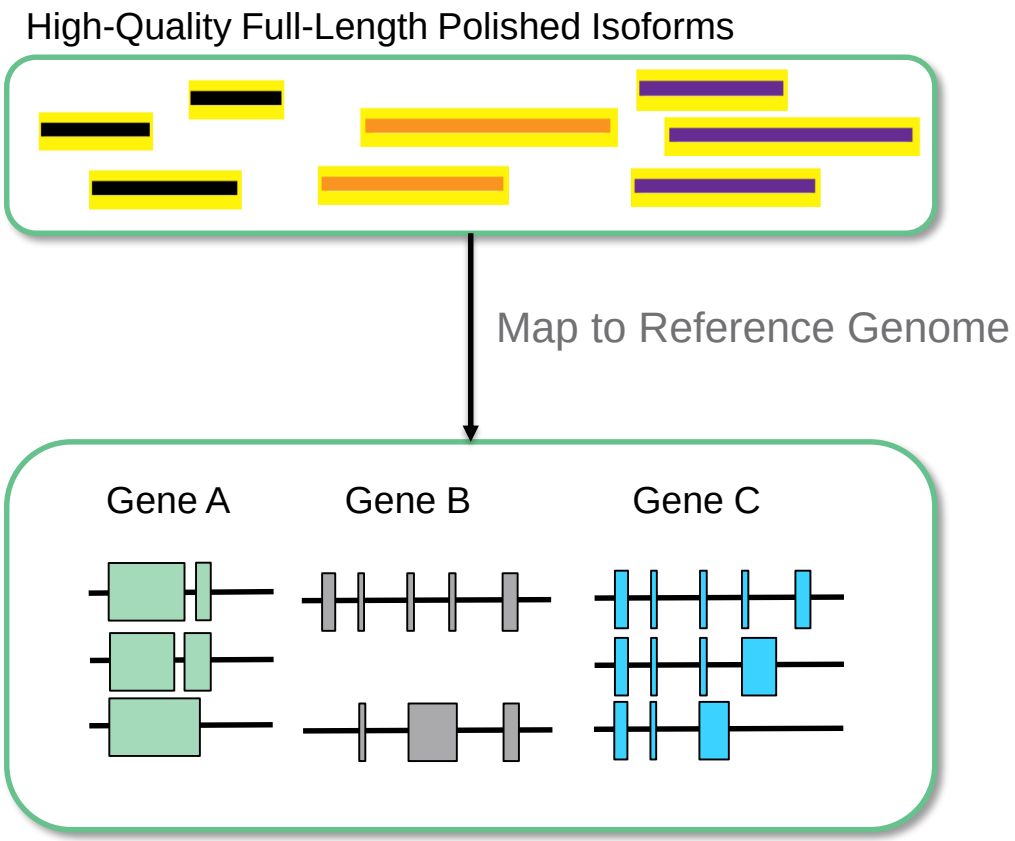
CCS → CLASSIFY FULL-LENGTH → CLUSTER ISOFORMS → MAP & COLLAPSE

Iso-Seq 1 and 2





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



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

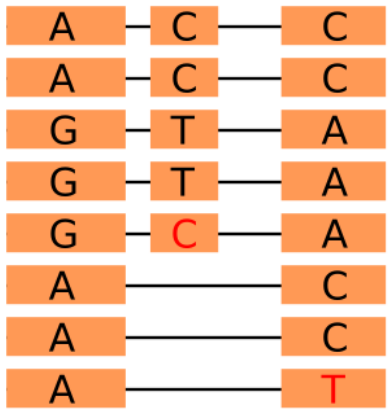
- Supported by SMRT Link
  - Use Iso-Seq analysis application in SMRT Link
  - Provide barcoded sequences as parameter to Classify step
- May use community script to get per barcode count information for each transcript after Iso-Seq is run

PACBIO®

Iso-Seq Community Tools

ISOPHASE: ISOFORM PHASING USING ISO-SEQ DATA

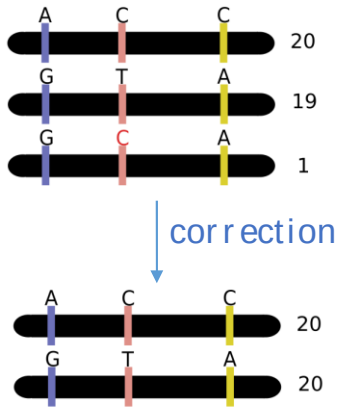
ALIGNMENT



SNP CALLING

Position	SNPs
POS1	A,G
POS2	C,T
POS3	C,A

PHASING



VCF OUTPUT

```
##fileformat=VCFv4.2
#CHROM  POS ID  REF ALT QUAL  FILTER  INFO            FORMAT  ISOFORM1   ISOFORM2
chr1    105 .   A   G   .     PASS    DP=40;AF=0.50   GT:HQ   0|1:20,20  0:15
chr1    190 .   C   T   .     PASS    DP=40;AF=0.50   GT:HQ   0|1:20,20  0:15
chr1    336 .   C   A   .     PASS    DP=40;AF=0.50   GT:HQ   0|1:20,20  0:15
```

ANGUS X BRAHMAN F1 CATTLE

Genome Assembly

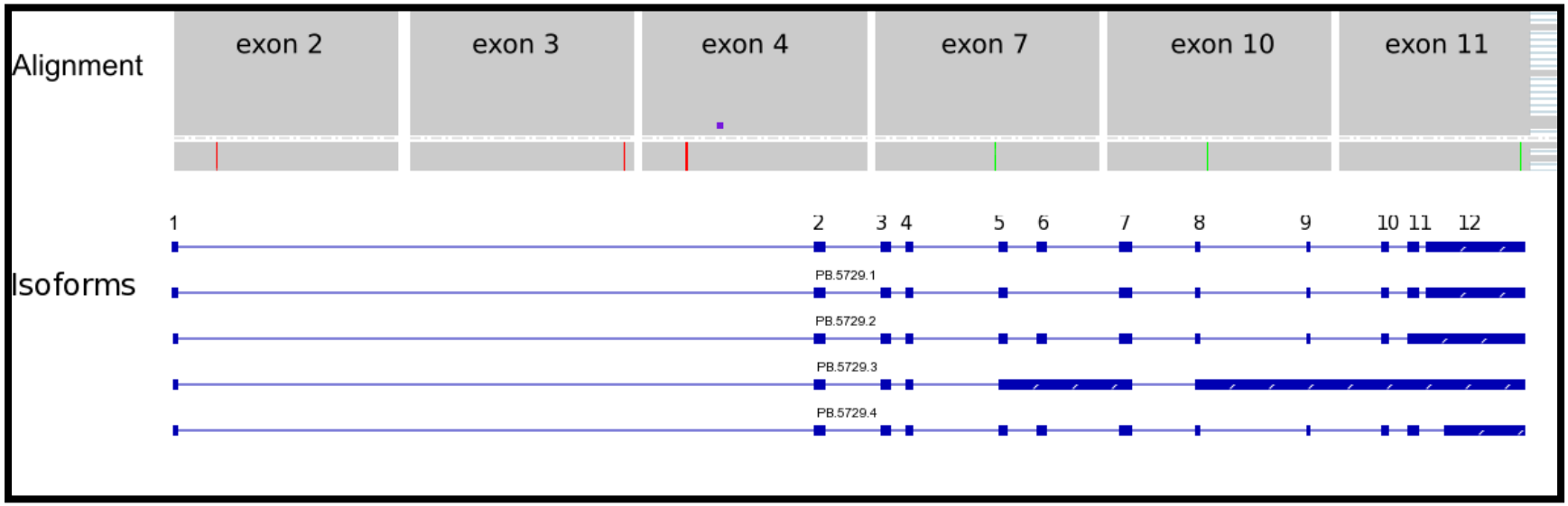
- Angus (sire) x Brahman (dam) F1 cattle
- 115-fold coverage on PacBio RS II and Sequel systems
- Assembled using Falcon
- ~90% of genome phased using Unzip

CONTIG	NUMBER	LENGTH	N50	LONGEST
PRIMARY	1427	2.71 Gb	31.4 Mb	65.3 Mb
HAPLOTIGS	5879	2.45 Gb	2.48 Mb	14.0 Mb

Iso-Seq Transcriptome Data

- 8 Sequel cells of tissues from single individual
- Analyzed using IsoSeq2
- Mapped to genome with ≥ 99% coverage, ≥ 95% identity
- 30,137 final isoforms (12,101 genes)
- Selected for phasing: 1758 genes with ≥ 40 full-length CCS read coverage

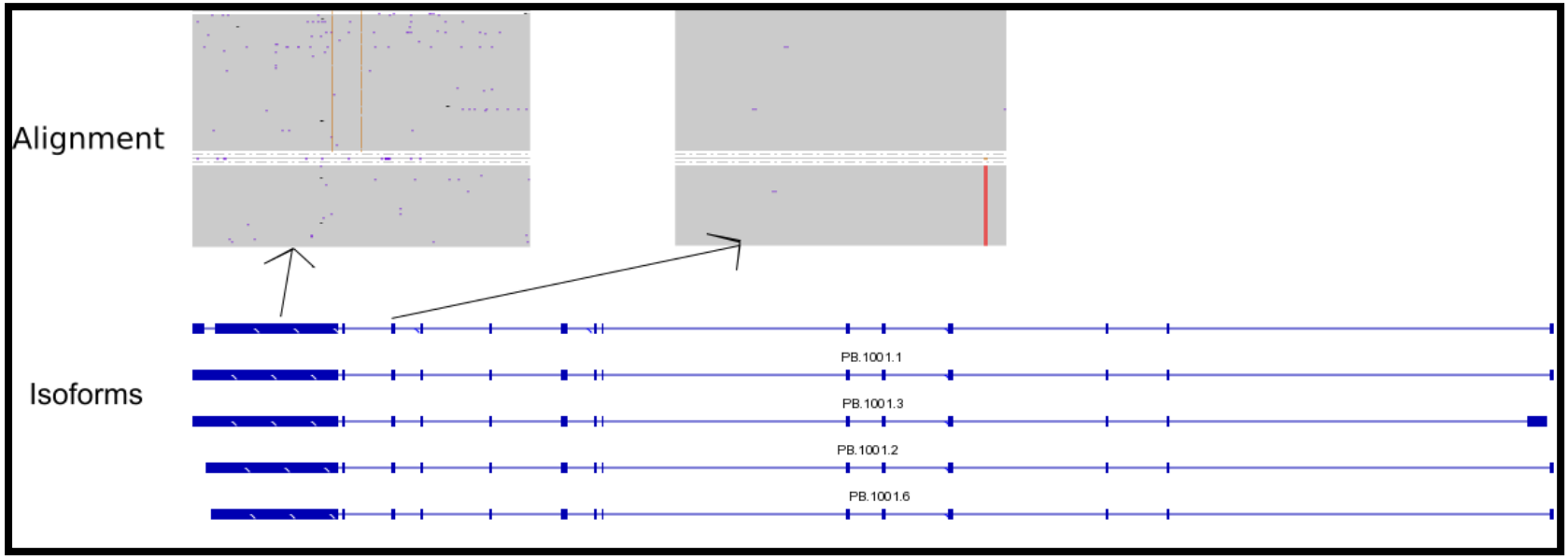
EXAMPLE OF SNP CALLING VERIFIED BY GENOME



There are 5 different isoforms for this gene. All isoforms cover all 6 SNP sites.

All 6 SNPs validated by genome assembly Unzip results

VPS36 ISOFORMS CALLED SNPS NOT PHASED IN GENOME

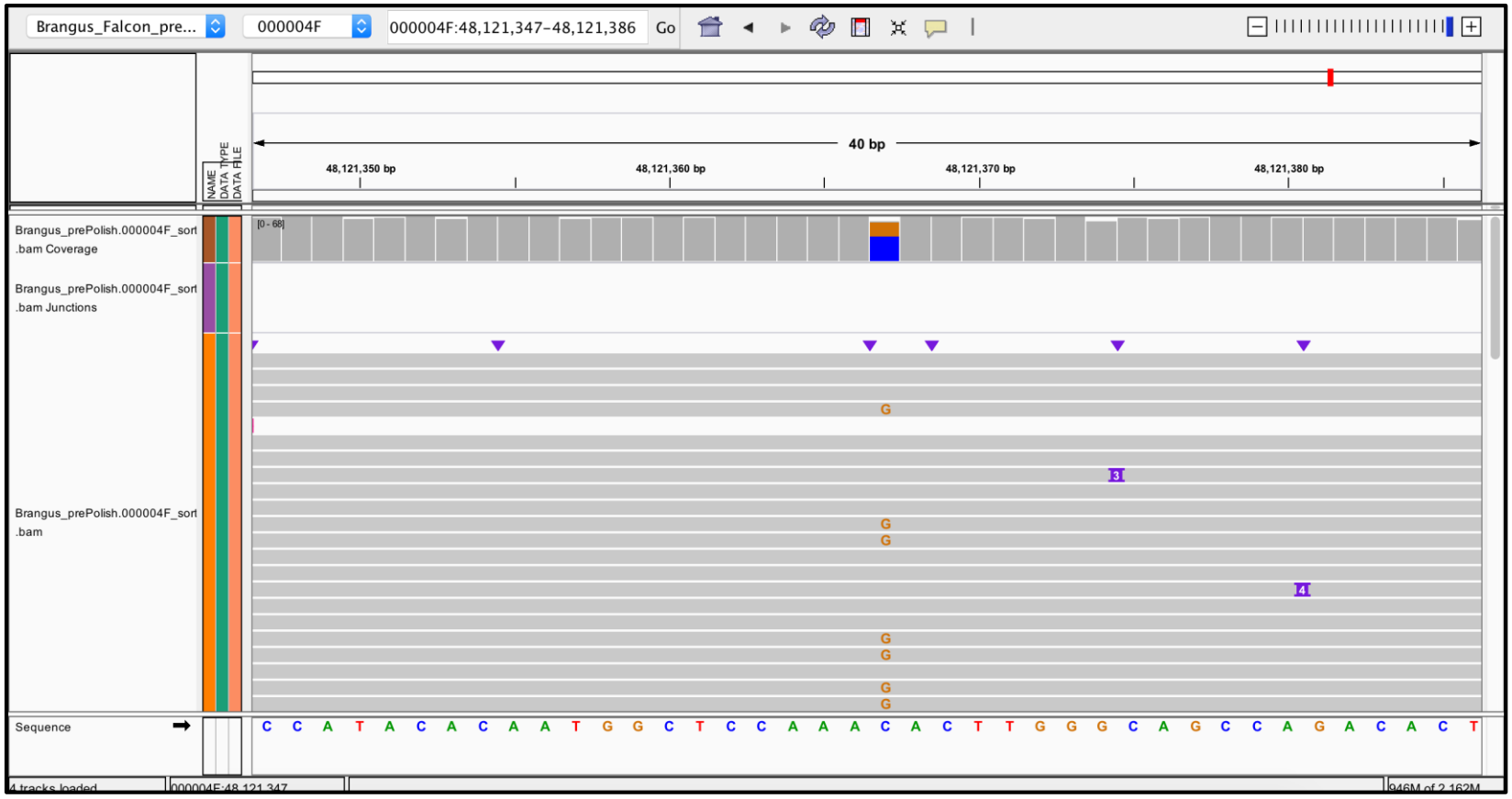


This gene (PB.1001, VPS36) contains 228 FL reads.

- Strong evidence for the 3 SNPs.
- Unzip did not phase this region – so, are the SNPs supported by genome?

VPS36 ISOFORMS CALLED SNPS NOT PHASED IN GENOME

The first SNP 000004F|arrow|arrow:48163477 (C->G) is supported in the pre-polish BAM file.

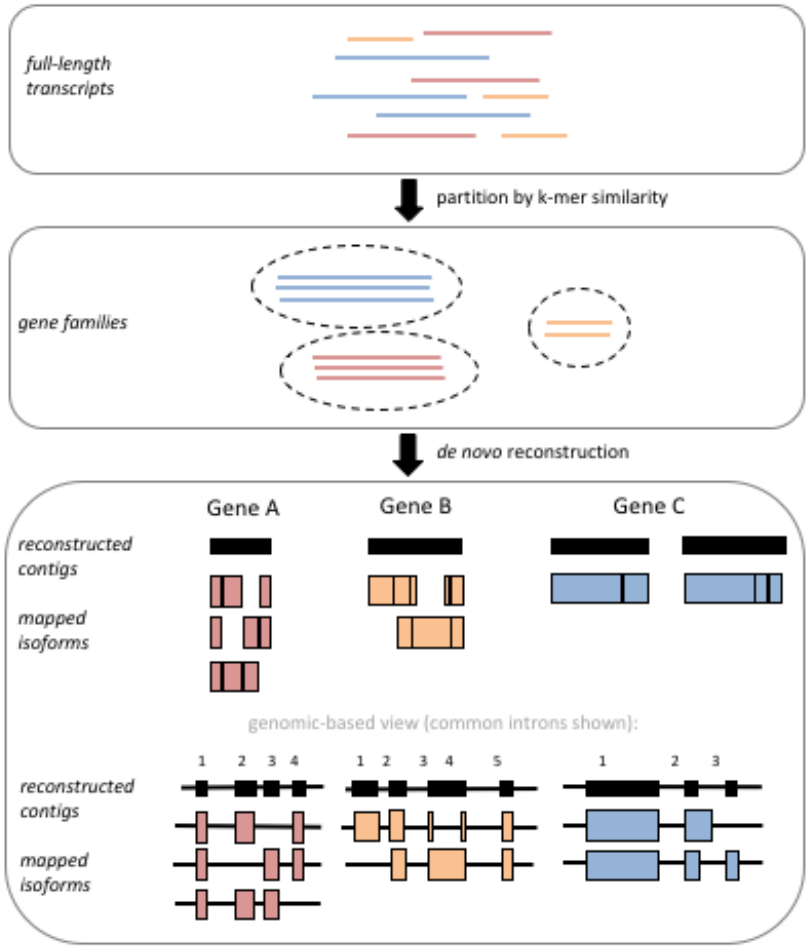


COGENT: RECONSTRUCT CODING REGION

- Cogent
- No or poor reference genome

- Input: Iso-Seq high-quality isoforms
- Output: 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)

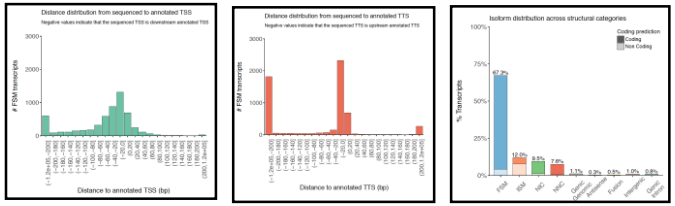
Iso-Seq Community Tools List

SQANTI & TAPPAS: QUALITY CONTROL, EVALUATION AND VISUALIZATION

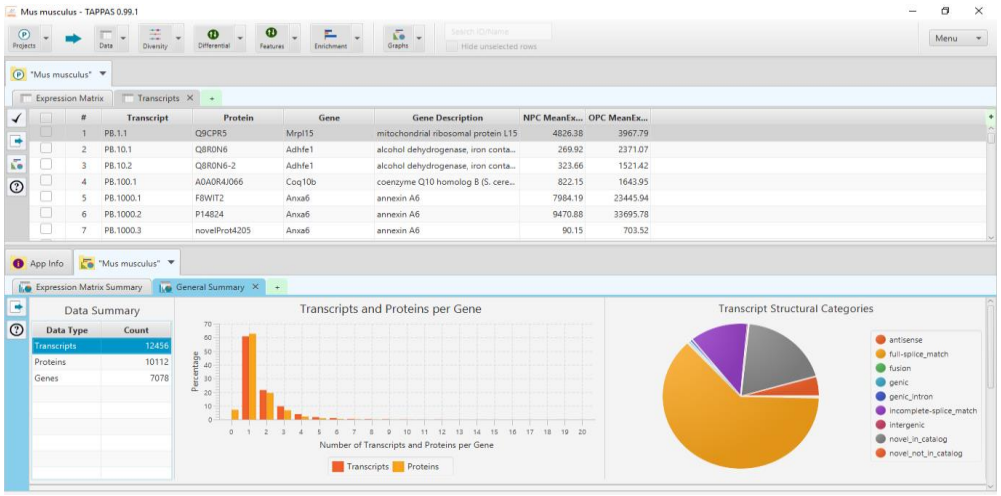
Developed by Ana Conesa Lab (U of FL)

SQANTI

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



TAPPAS visualize data at isoform level



www.pacb.com

PACBIO®

Iso-Seq3

Ultra Fast + High Performance + Scalable

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-SEQ 3 OVERVIEW

Iso-Seq 3 workflow

- CCS - same

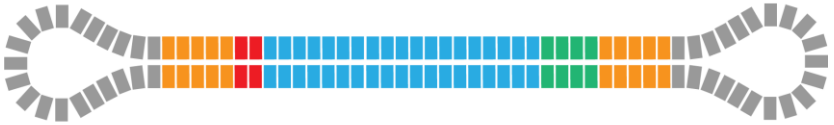
ISO-SEQ 3 OVERVIEW

Iso-Seq 3 workflow

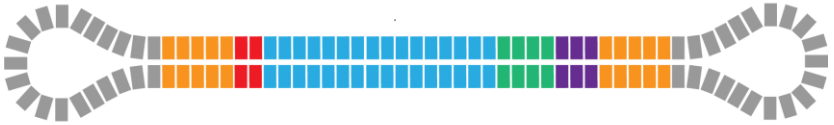
- CCS – same
- Classify – utilizing demultiplex barcoding algorithm (LIMA) with special `--isoseq` mode

ISO-SEQ 3 CLASSIFY SUPPORTS DIFFERENT LIBRARY PREP

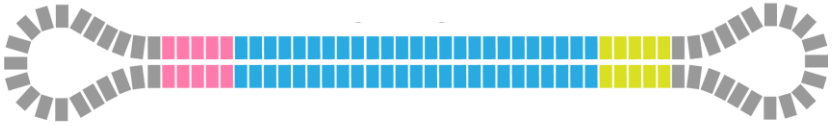
Whole transcriptome

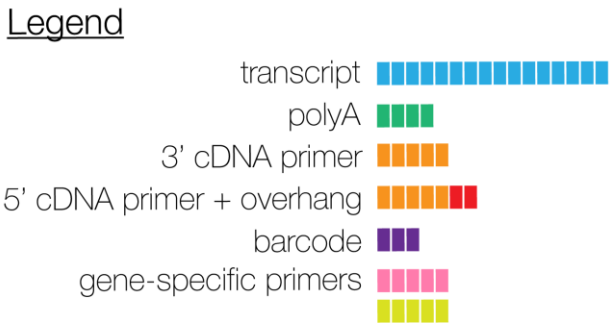


Whole transcriptome, barcoded



Targeted genes





ISO-SEQ 3 CLASSIFY REMOVES MORE ARTIFACTS

Full Length:

TSO Artifact:

ISO-SEQ 3 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 by detecting additional cDNA primers in the middle of the sequence

LIBRARY ARTIFACTS

TYPE	CAUSE	ISO-SEQ 1	ISO-SEQ 3
TSO Artifacts	Template switching artifacts	no	yes
Artificial Concatemers	Insufficient SMRT adapter	yes	yes

- Iso-Seq 3 removes TSO artifacts are part of the demultiplexing (lima) process
- Iso-Seq 3 removes concatemers as the first part of the Cluster step (see later slides)

ISO-SEQ 3 WORKFLOW

Iso-Seq3 workflow

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

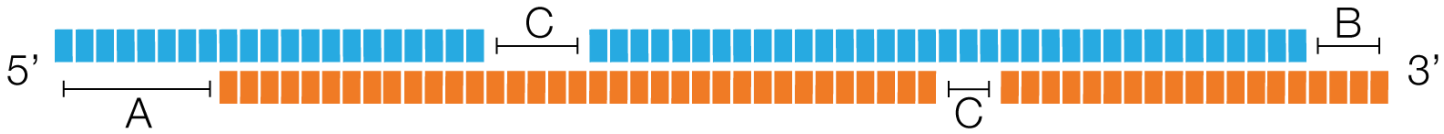
ISO-SEQ 3 CLUSTER: ISOFORM DEFINITION

Two Full-Length reads are considered the same isoform if they are:

(A) <100 bp difference in 5’ start

(B) <30 bp difference in 3’ end

(C) <10 bp in internal gap (exon), no limit on the number of gaps



ISO-SEQ 3: 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% AND ≥2 FL read support

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-SEQ 1 VS ISO-SEQ 3

Best practice for analyzing multiplexed Iso-Seq data

	Iso-Seq 1	Iso-Seq 3
Runtime	~3 days for 3 cells	~14 hr for 3 cells
Memory usage	High	Low
Library artifact detection	Poor	Good
Demultiplexing accuracy	OK	Good
Can analyze by multiplexed barcode?	GUI and command line	command line-only

- Iso-Seq 1 will be removed in future releases. With few exceptions, customers will find Iso-Seq 3 or Iso-Seq 3 with Mapping to be suited for their needs.

DEMULTIPLEXING DATA AFTER ISO-SEQ ANALYSIS

GitHub Tutorial: Demultiplexing SMRT Link Iso Seq Jobs

Tutorial: Demultiplexing SMRT Link Iso Seq Jobs

Elizabeth Tseng edited this page 2 minutes ago · 1 revision

Last Updated: 08/21/2018

This tutorial is for demultiplexing Iso-Seq 1 and Iso-Seq 3 jobs in SMRT Link 6.0. This script will also work with developers version of Iso-Seq3 following this tutorial. The scripts offered below are standalone --- no installation required, however, you will need to have Python and also BioPython library installed.

- Identifying the Unix path for your SMRT Link job
- Demultiplexing Iso-Seq 1 and Iso-Seq 3 jobs without a reference genome
- Demultiplexing Iso-Seq 1 and Iso-Seq 3 jobs with a reference genome

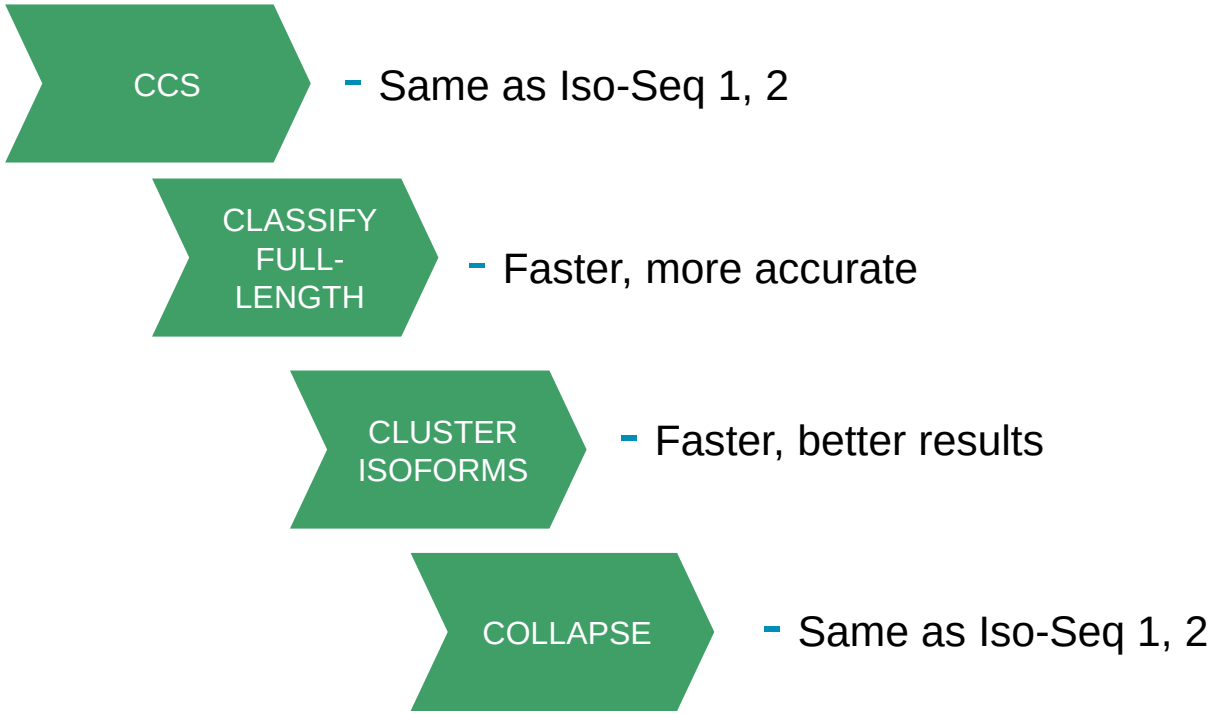
Prerequisite

- Python (2.7.x)
- BioPython

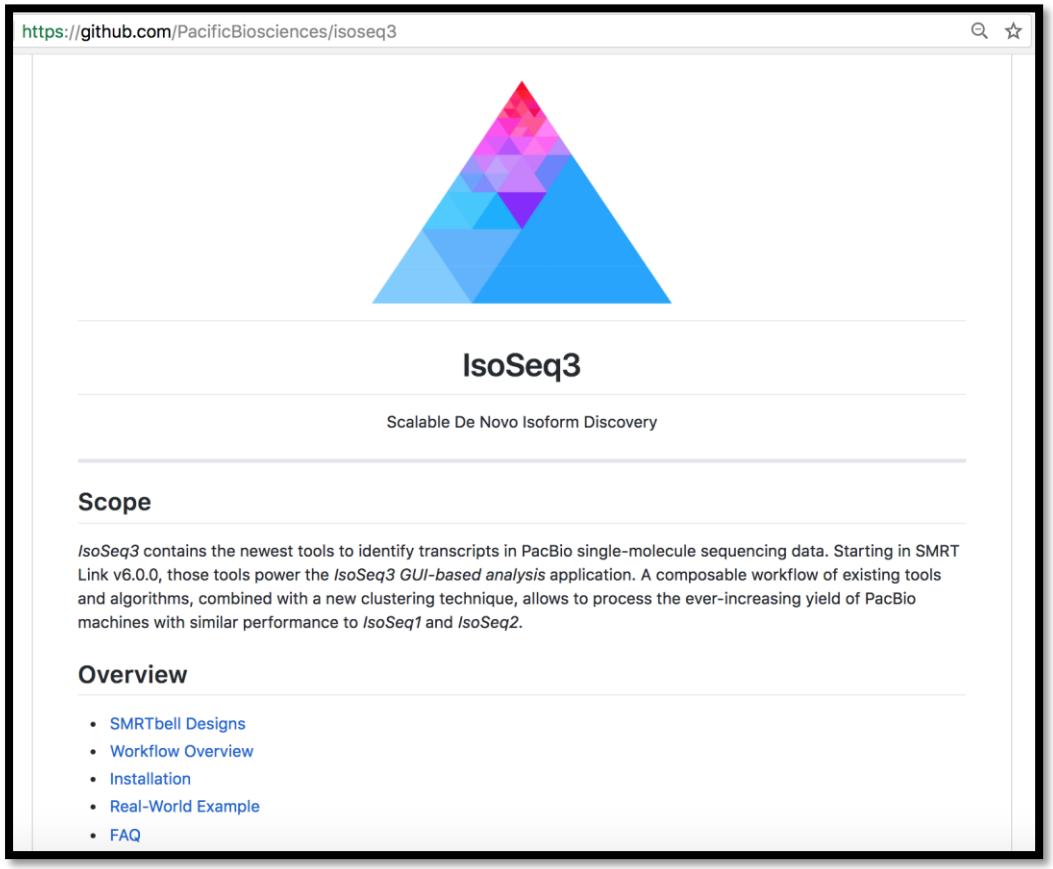
How to use the scripts

No installation is required. You can directly copy the scripts to your local folder:

ISO-SEQ 3 IMPROVEMENTS COMPARED TO ISO-SEQ 1&2



- 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 (due out in October)

www.pacb.com

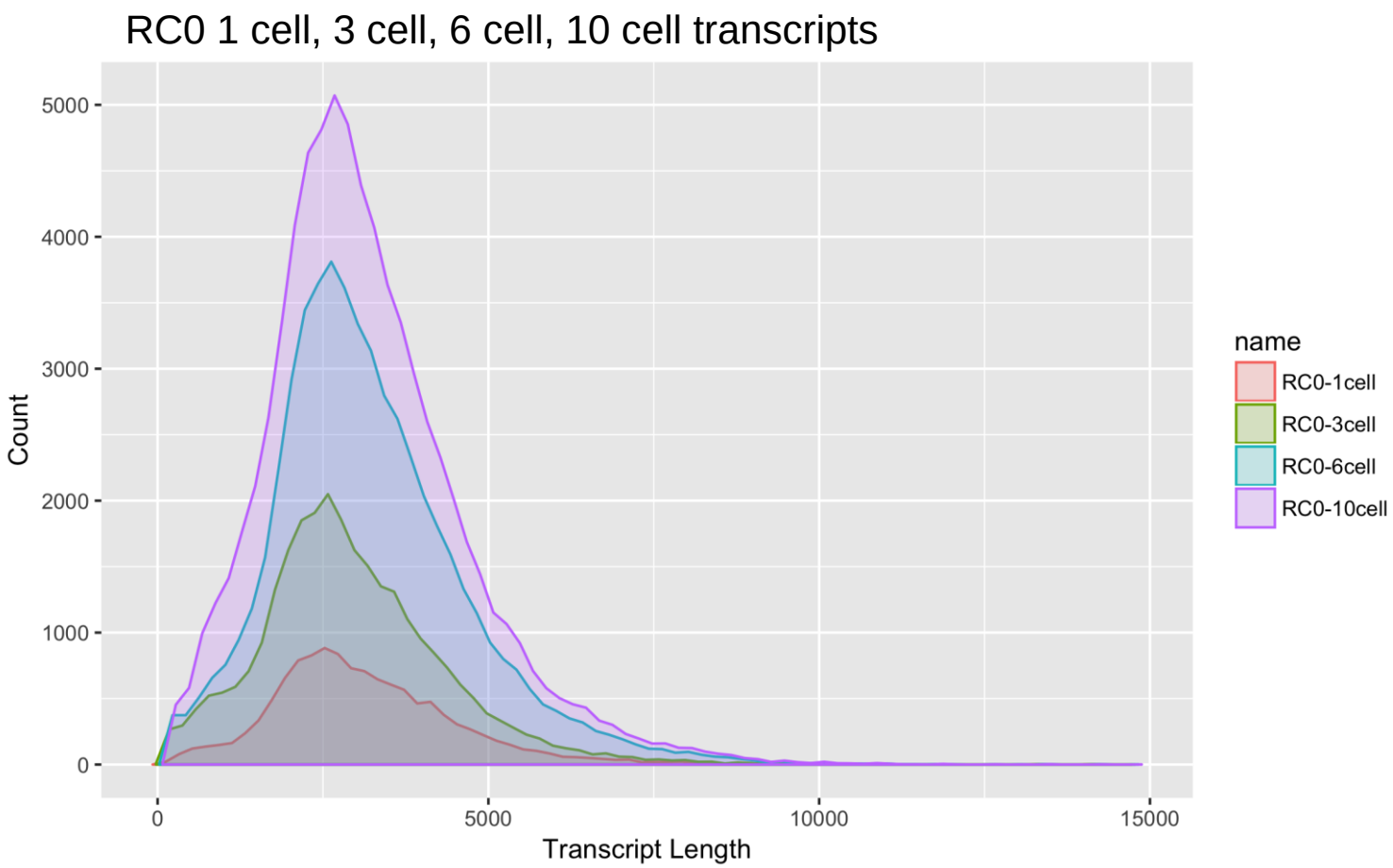
Iso-Seq3 Performance

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

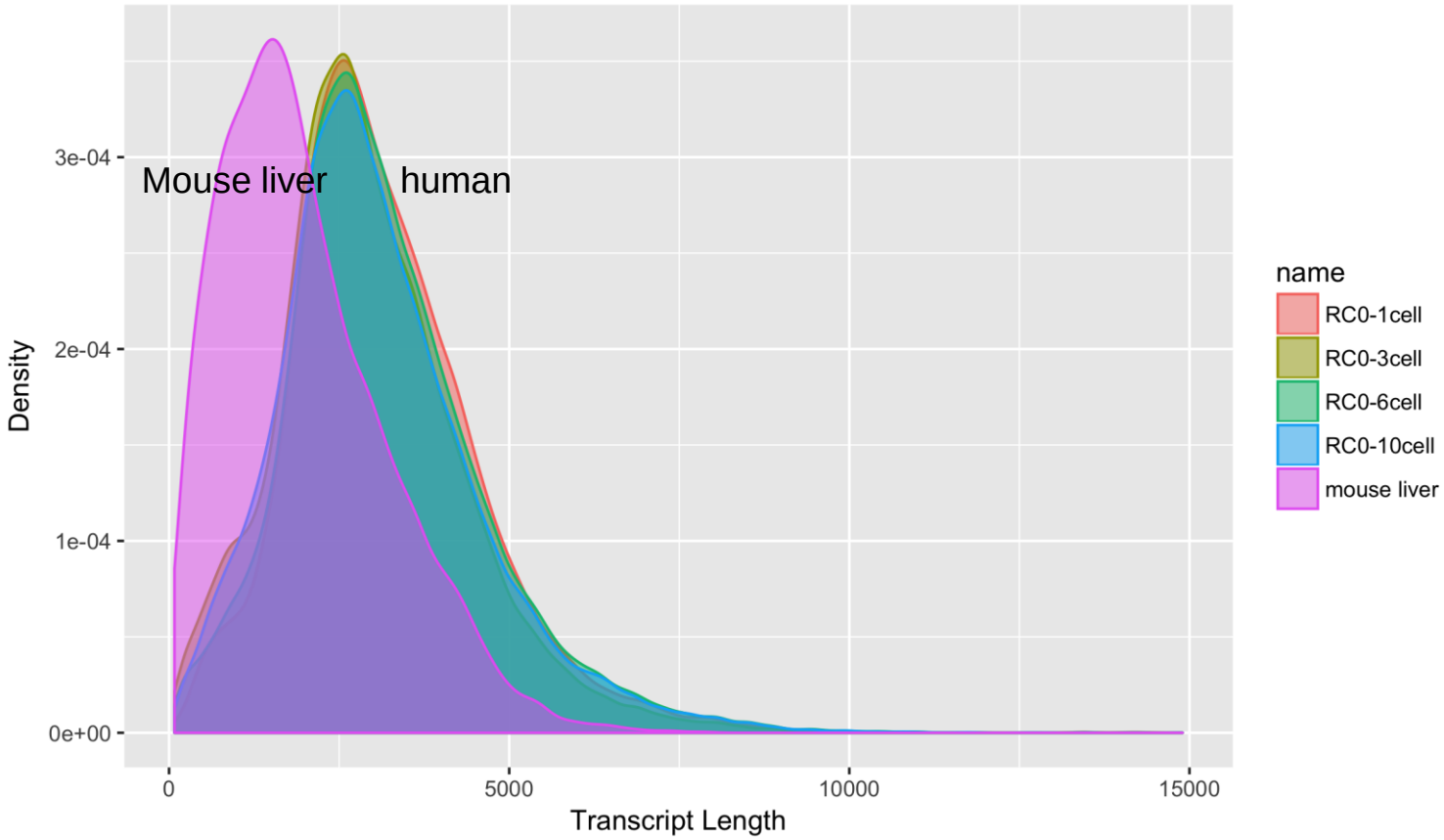
- RC0 = Universal Human Reference RNA (human) + Lexogen SIRV spike-in controls
- Not including CCS and Mapping runtime
- Computing configuration：16 CPU / node
- Tested using command line

HUMAN TRANSCRIPTS LENGTH DISTRIBUTION

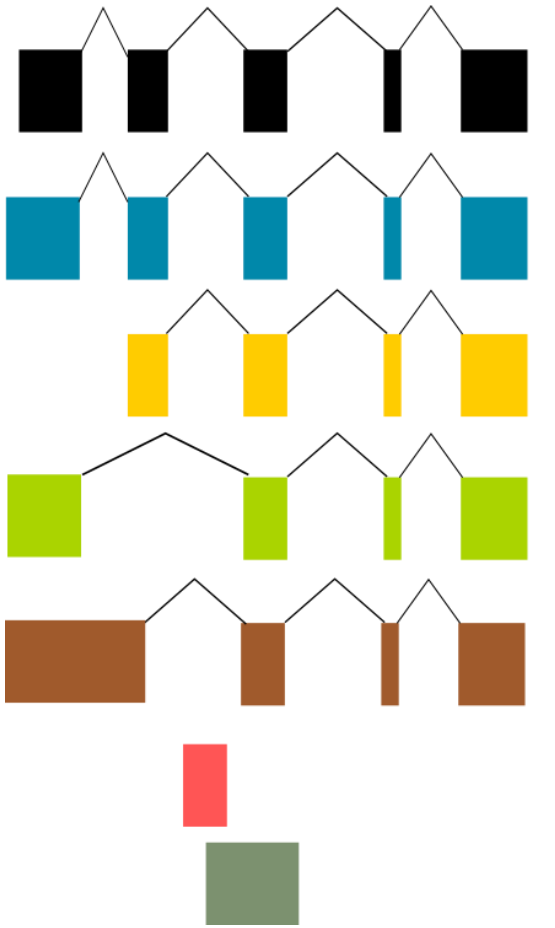


DIFFERENCE BETWEEN HUMAN AND MOUSE LIVER TRANSCRIPTS

Mouse liver transcripts slightly shorter than RC0



USE SQANTI* TO EVALUATE ISO-SEQ3 RESULTS



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 junction

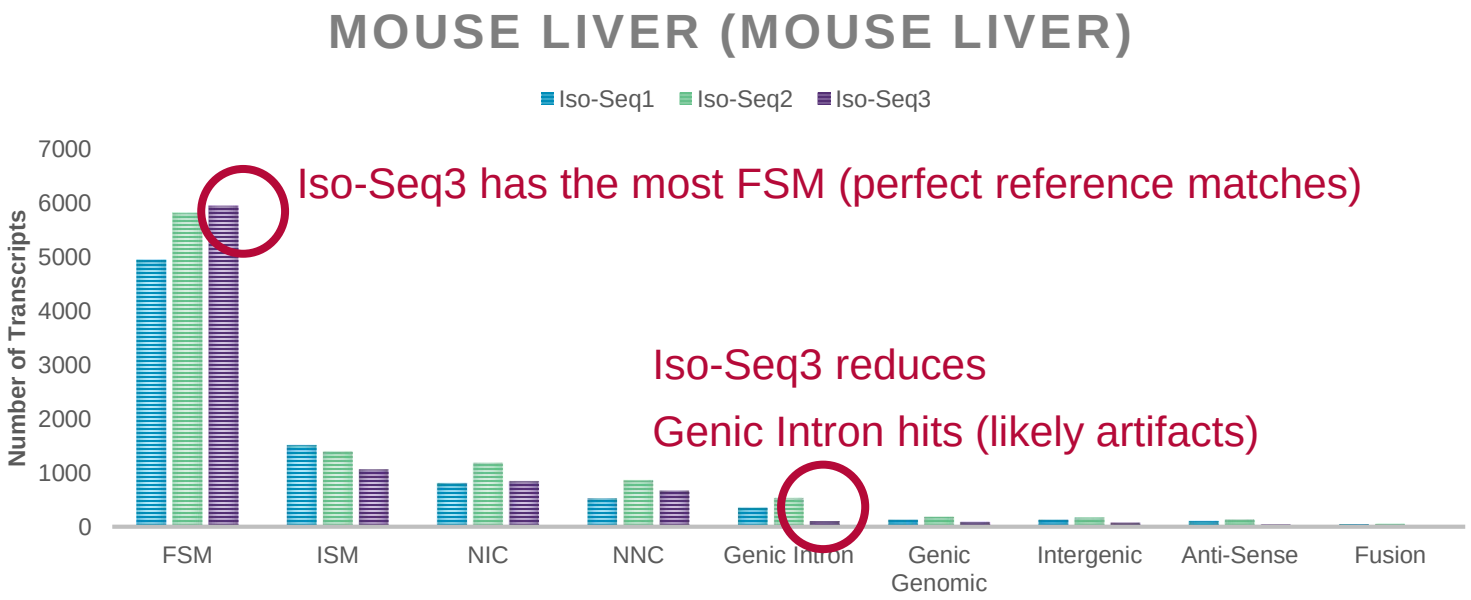
Genic Intron — within intron

Genic Genomic — Overlap with intron and exons

*SQANTI is a community tool developed by Conesa lab

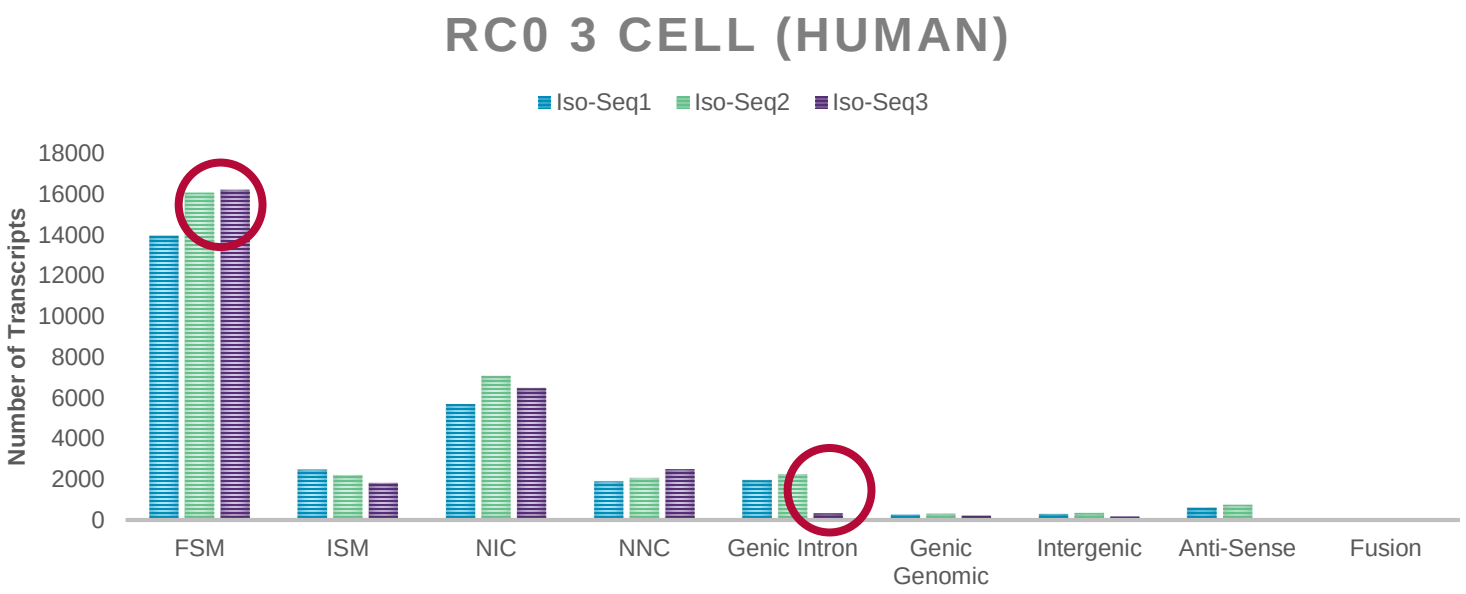
Tardaguila, M. *et al.* SQANTI: extensive characterization of long read transcript sequences for quality control in full-length transcriptome identification and quantification. 1–31 (2017). doi:10.1101/118083

ISO-SEQ3 VS REF ANNOTATION: MOUSE LIVER



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

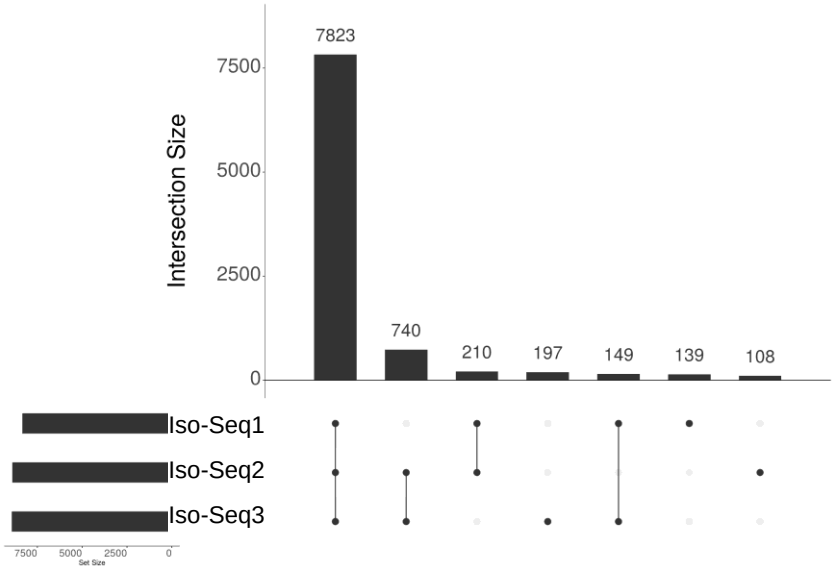
ISO-SEQ3 VS REF ANNOTATION: HUMAN



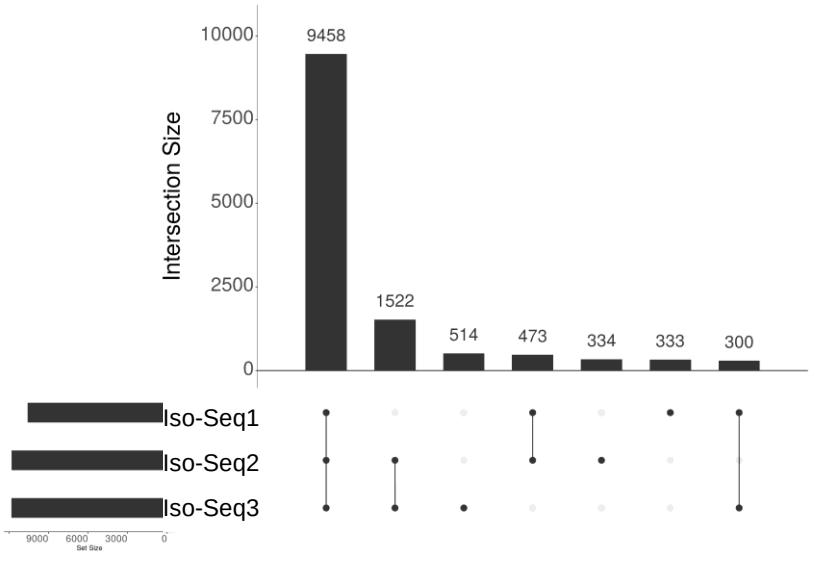
SQANTI : compare Iso-Seq results vs Gencode v27 Reference Gene Annotation

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

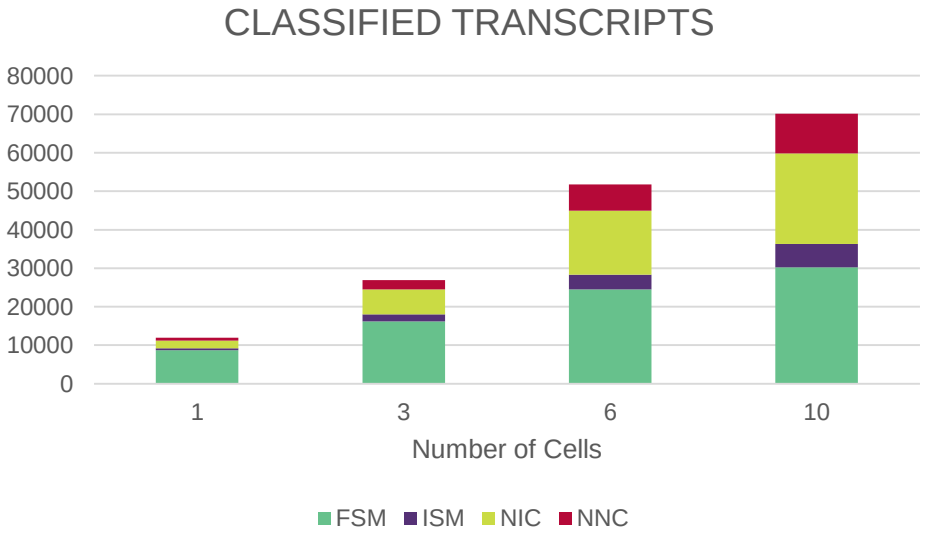
HOW MUCH SEQUENCING IS NEEDED?

FSM = Full Splice Match

ISM = Incomplete Splice Matches

NIC = Novel In Catalog

NNC = Novel Not in Catalog



ISO-SEQ BIOINFORMATICS BREAKOUT SUMMARY

1) Iso-seq is a robust method for characterizing transcriptional diversity
2) The Iso-Seq3 informatics pipeline provides a streamlined workflow for baseline identification of unique isoforms
3) Community tools built around the Iso-Seq workflow enhance your ability to profile transcriptional activity in PacBio data

COMMUNITY SUPPORT FOR ISO-SEQ USERS AND DEVELOPERS

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

www.pacb.com