PACIFIC BIOSCIENCES®

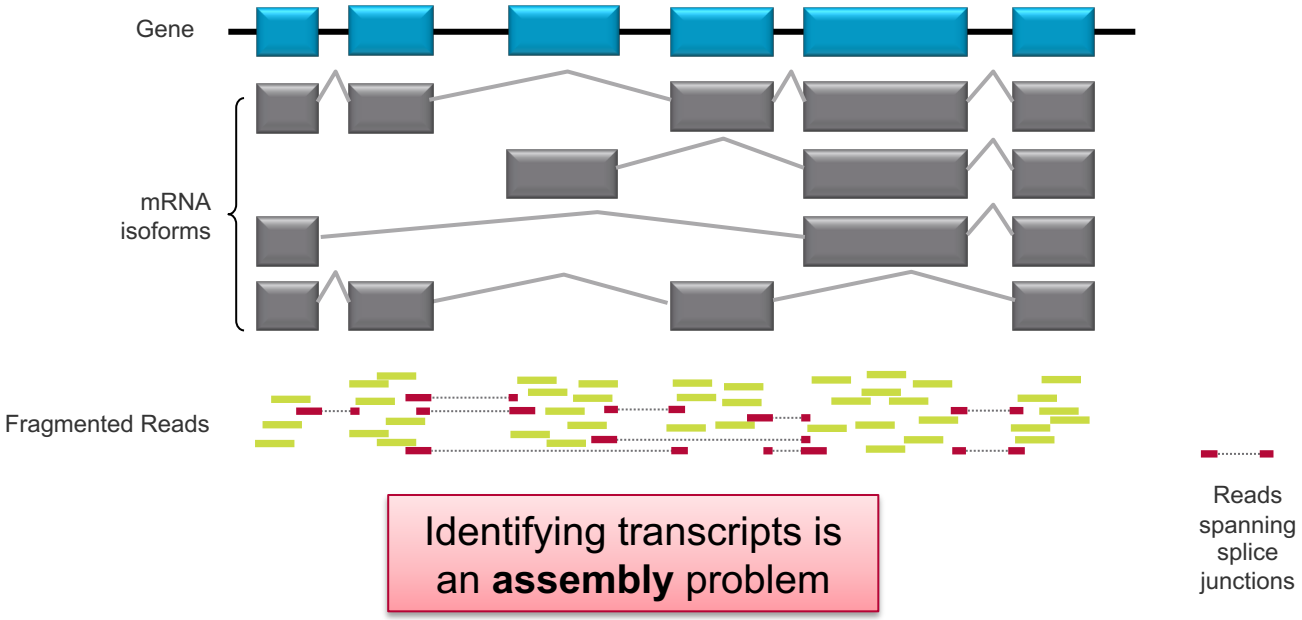
Bulk and Single-Cell Isoform Sequencing Using PacBio Long Reads

Elizabeth Tseng, Associate Director, Product Marketing, PacBio

@Magdoll

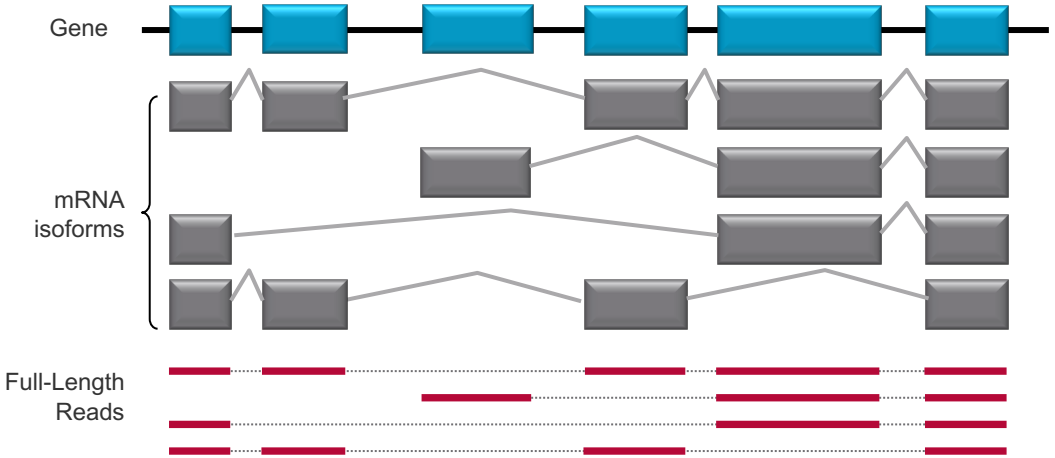
What is Iso-Seq method?

RNA-SEQ CANNOT RESOLVE COMPLEX SPLICING



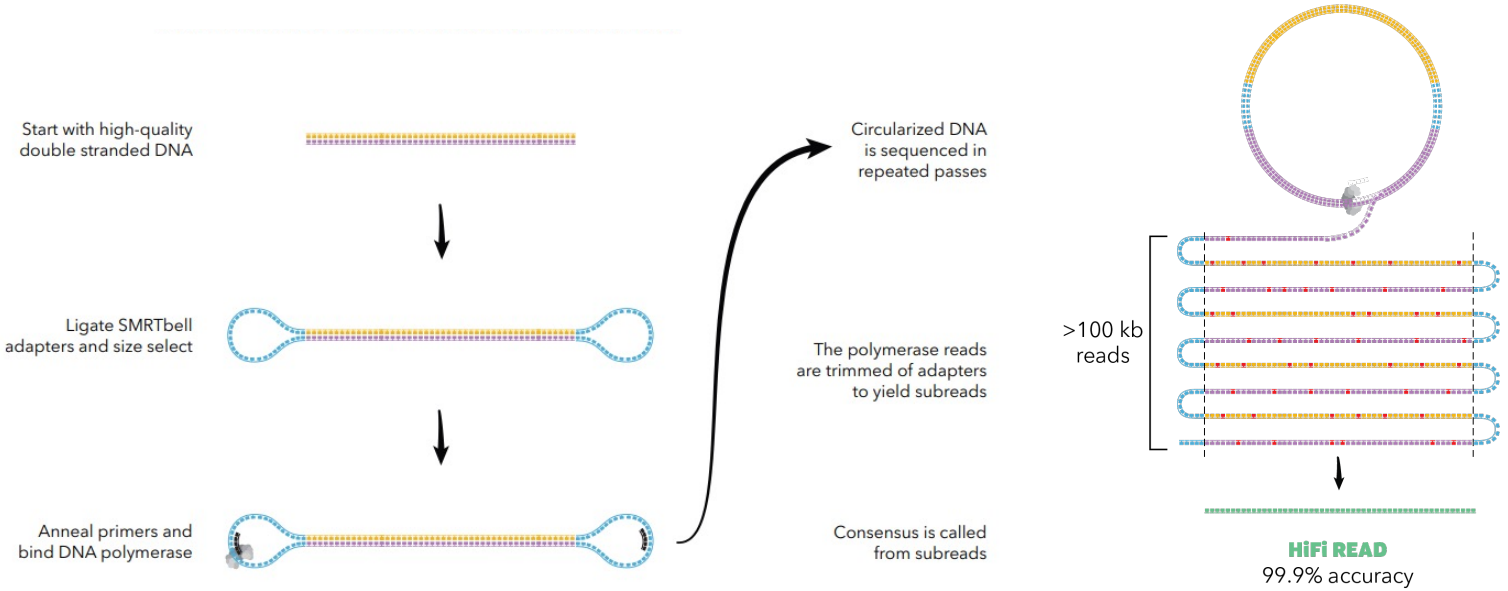
Identifying transcripts is an **assembly** problem

ISO-SEQ METHOD: FULL-LENGTH TRANSCRIPT SEQUENCING



No assembly required

HIFI READS PROVIDE ACCURACY FOR ISO-SEQ ANALYSIS

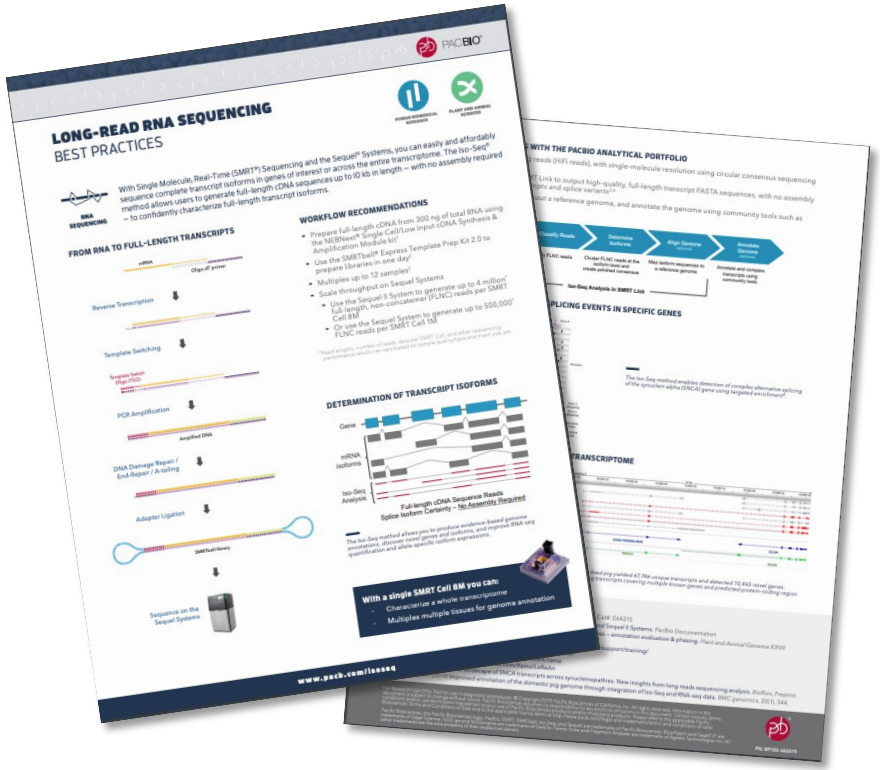


ISO-SEQ WORKFLOW END-TO-END SOLUTION

LIBRARY PREP
1 DAY

SMRT SEQUENCING
1 DAY

DATA ANALYSIS
1 DAY



www.pacb.com/iso-seq

ISO-SEQ WORKFLOW END-TO-END SOLUTION

LIBRARY PREP

1 DAY

Iso-Seq Express kit

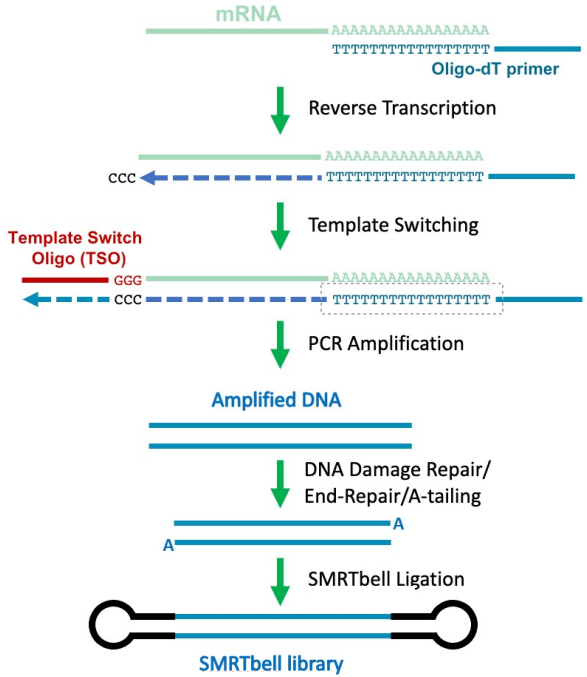
- 60-300 ng total RNA
- Full-length cDNA
- Multiplexing support

SMRT SEQUENCING

1 DAY

Sequel II System

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

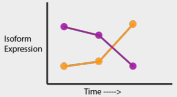


https://www.pacb.com/wp-content/uploads/Procedure-Checklist-Iso-Seq-Express-Template-Preparation-for-Sequel-and-Sequel-II-Systems.pdf

ISO-SEQ WORKFLOW END-TO-END SOLUTION

DATA ANALYSIS

1 DAY

		Input	Output
Sequencing			ccs.bam
Iso-Seq Analysis		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	

Supporting Tools

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

- gene family finding
- genome reconstruction
- evaluate assembly

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

TALON

- long read processing & annotation pipeline developed independently by ENCODE4

ISO-SEQ WORKFLOWS

Genome Annotation

Sequencing 

Iso-Seq Analysis 

Transcript Classification 

or MAKER

AUGUSTUS

Functional Annotation 

Differential Expression

Sequencing 

Iso-Seq Analysis 

Transcript Classification 

Functional Annotation 

Differential Analysis 

or DESeq

(+RNA-Seq)

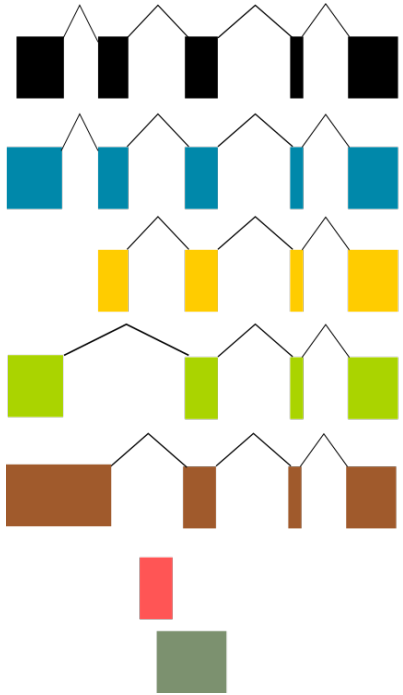
Phasing

Sequencing 

Iso-Seq Analysis 

IsoPhase

SQANTI(3): A New Classification Of Transcripts



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

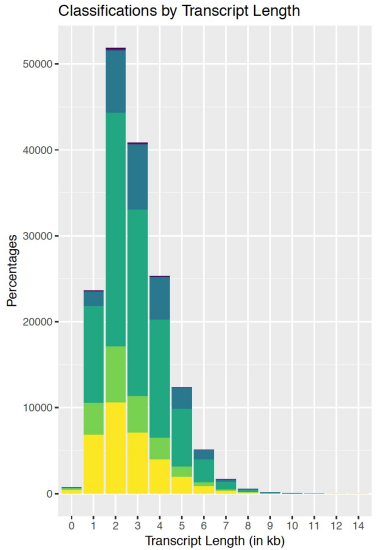
Genic Intron Within intron

Genic Genomic Overlap with introns 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-SEQ METHOD ON THE SEQUEL II SYSTEM

- **Comprehensive**
- Full-Length
- Highly accurate



162,290 transcripts

80 – 14,288 bp
(mean: 3.3 kb)

Dataset: Alzheimer brain on 1 SMRT Cell 8M

ISO-SEQ METHOD ON THE SEQUEL II SYSTEM

- **Comprehensive**
- Full-Length
- Highly accurate

	Known	Novel	Total
Genes	17,051	619	**17,670**
Isoforms	51,660	110,630	**162,290**

Dataset: Alzheimer brain on 1 SMRT Cell 8M

ISO-SEQ METHOD ON THE SEQUEL II SYSTEM

- Comprehensive
- **Full-Length**
- Highly accurate

Category	Count	Description
FSM	32,649	Perfect match
ISM	19,011	Incomplete match
NIC	84,610	Novel isoform using known junctions
NNC	25,323	Novel isoform using at least novel junction
Antisense	321	Anti-sense to known gene
Intergenic	376	Intergenic

Dataset: Alzheimer brain on 1 SMRT Cell 8M

ISO-SEQ METHOD ON THE SEQUEL II SYSTEM

- Comprehensive
- **Full-Length**
- Highly accurate

Category	Count	CAGE peak within 50 bp
FSM	32,649	70%
ISM	19,011	37%
NIC	84,610	36%
NNC	25,323	57%
Antisense	321	24%
Intergenic	376	24%

Dataset: Alzheimer brain on 1 SMRT Cell 8M

ISO-SEQ METHOD ON THE SEQUEL II SYSTEM

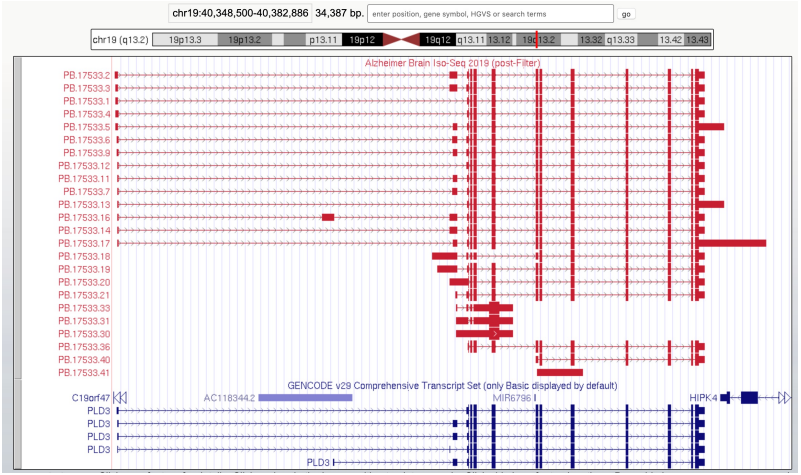
- Comprehensive
- **Full-Length**
- Highly accurate

Category	Count	CAGE peak within 50 bp	polyA Motif Detected
FSM	32,649	70%	72%
ISM	19,011	37%	62%
NIC	84,610	36%	55%
NNC	25,323	57%	72%
Antisense	321	24%	43%
Intergenic	376	24%	38%

Dataset: Alzheimer brain on 1 SMRT Cell 8M

ISO-SEQ METHOD ON THE SEQUEL II SYSTEM

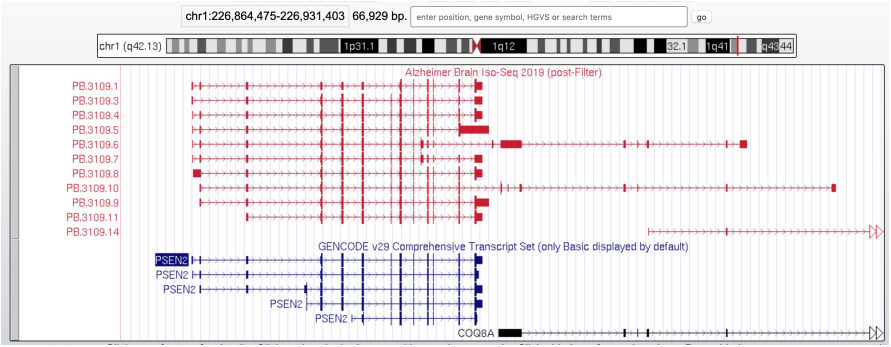
- Comprehensive
- **Full-Length**
- Highly accurate



Dataset: Alzheimer brain on 1 SMRT Cell 8M

ISO-SEQ METHOD ON THE SEQUEL II SYSTEM

- Comprehensive
- **Full-Length**
- Highly accurate

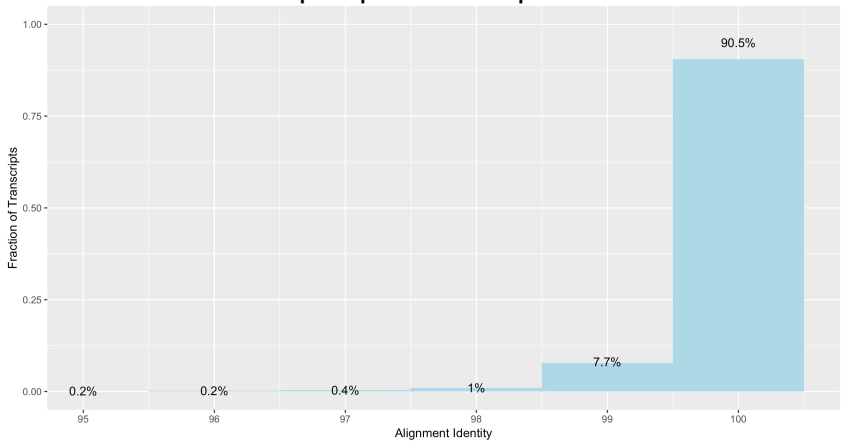


Dataset: Alzheimer brain on 1 SMRT Cell 8M

ISO-SEQ METHOD ON THE SEQUEL II SYSTEM

- Comprehensive
- Full-Length
- **Highly accurate**

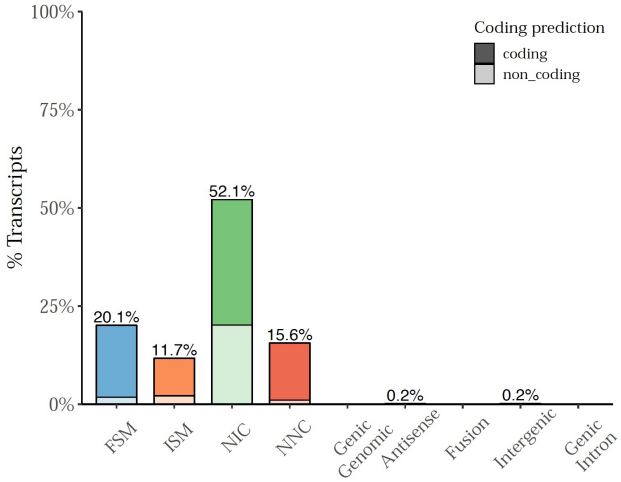
> 99% of Iso-Seq output transcript is >99% accurate



Dataset: Alzheimer brain on 1 SMRT Cell 8M

ISO-SEQ METHOD ON THE SEQUEL II SYSTEM

- Comprehensive
- Full-Length
- **Highly accurate**



Dataset: Alzheimer brain on 1 SMRT Cell 8M

Who is using Iso-Seq Method?

PACBIO®

Genome Annotation and Differential Expression Analysis using the Iso-Seq Method

DE NOVO TRANSCRIPTOME AND COMPARATIVE ANALYSIS IN RICE

International Journal of *Molecular Sciences*

MDPI

Article

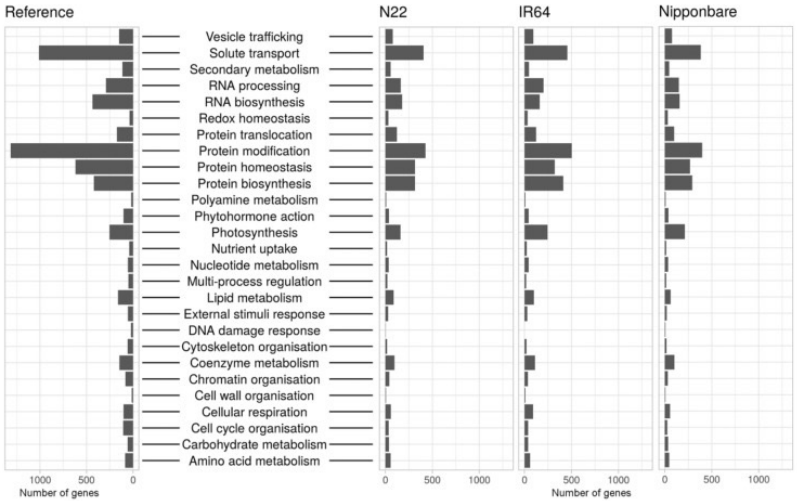
Utilizing PacBio Iso-Seq for Novel Transcript and Gene Discovery of Abiotic Stress Responses in *Oryza sativa* L.

Stephanie Schaarschmidt [1,*], Axel Fischer [1], Lovely Mae F. Lawas [1,2], Rejbana Alam [3], Endang M. Septiningsih [4], Julia Bailey-Serres [3], S. V. Krishna Jagadish [5,6], Bruno Huettel [7], Dirk K. Hincha [1,†] and Ellen Zuther [1,*]

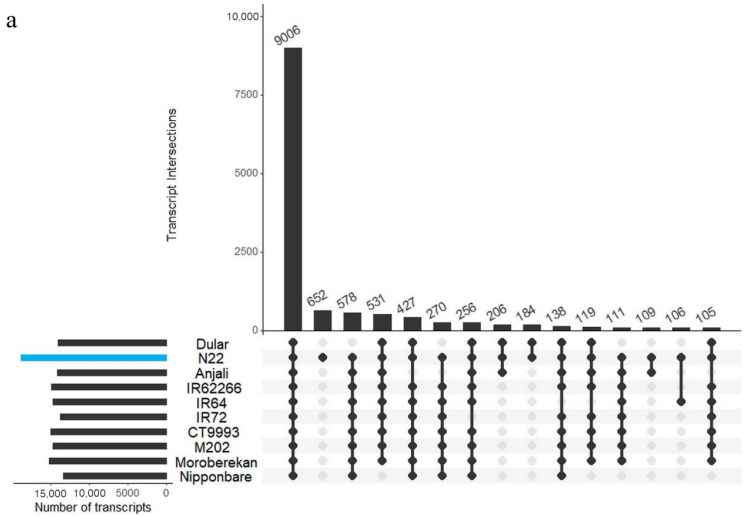
- Used the Iso-Seq method to build reference transcriptomes for 10 rice cultivates from multiple tissues
- Created a unique (collapsed) transcriptome for each cultivar combining both genome-mapped and unaligned Iso-Seq transcripts

DE NOVO TRANSCRIPTOME AND COMPARATIVE ANALYSIS IN RICE

Functional annotation across cultivars



Comparative analysis showed shared vs unique transcripts

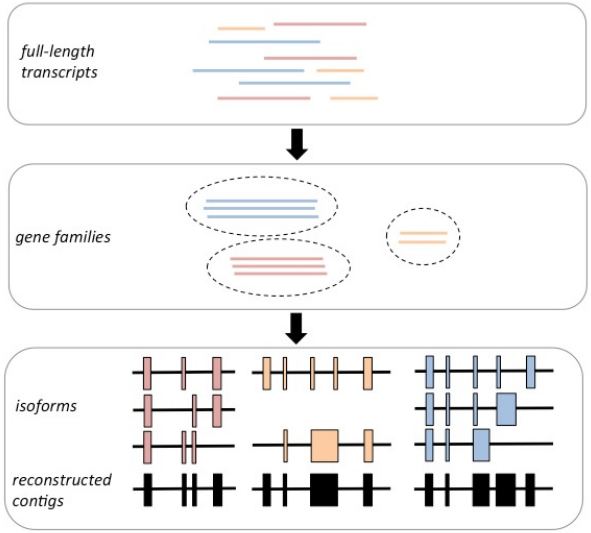


Schaarschmidt et al “Utilizing PacBio Iso-Seq for Novel Transcript and Gene Discovery of Abiotic Stress Responses in Oryza Sativa L.” *International Journal of Molecular Sciences* 21 (2021)

NO GENOME? NO PROBLEM!

COGENT workflow

Using only Iso-Seq data to find gene families and reconstruct a fake “genome”

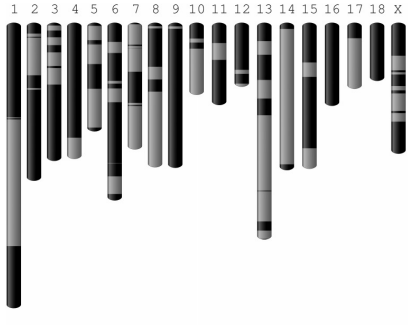


https://github.com/Magdoll/Cogent

Use COGENT results to…

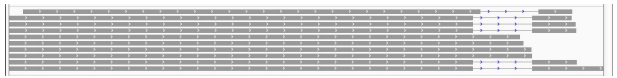
Evaluate genome assemblies

Pig Iso-Seq Cogent rescued 5 missing genes for the new pig assembly



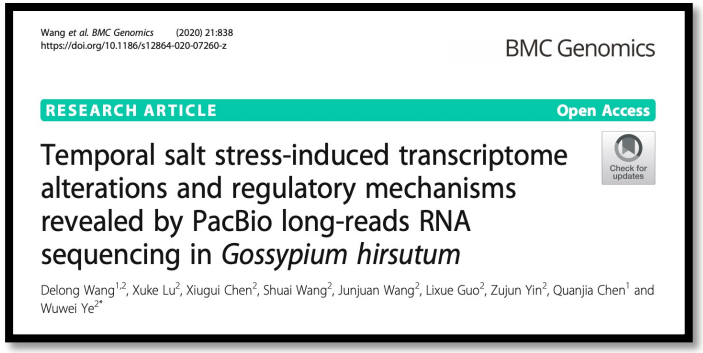
Visualize alternative splicing

You can still see skipped exons!



Warr, A. et al. (2020). “An Improved Pig Reference Genome Sequence to Enable Pig Genetics and Genomics Research.” *GigaScience*

OTHER WAYS FOR QUANTIFICATION



Wang *et al. BMC Genomics* (2020) 21:838
https://doi.org/10.1186/s12864-020-07260-z

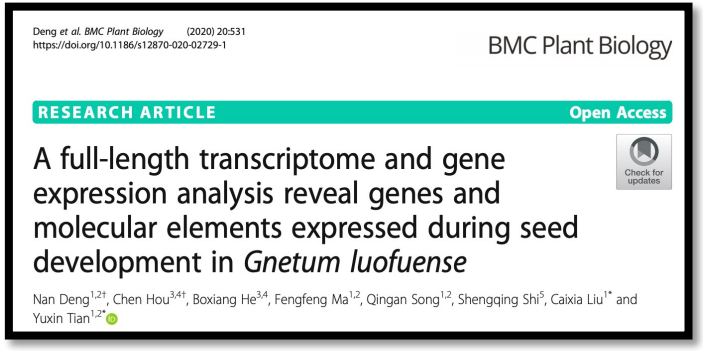
BMC Genomics

RESEARCH ARTICLE Open Access

Temporal salt stress-induced transcriptome alterations and regulatory mechanisms revealed by PacBio long-reads RNA sequencing in *Gossypium hirsutum*

Delong Wang[1,2], Xuke Lu[2], Xiugui Chen[2], Shuai Wang[2], Junjuan Wang[2], Lixue Guo[2], Zujun Yin[2], Quanjia Chen[1] and Wuwei Ye[2*]

Using UMI-tagged Iso-Seq reads directly for quantification



Deng *et al. BMC Plant Biology* (2020) 20:531
https://doi.org/10.1186/s12870-020-02729-1

BMC Plant Biology

RESEARCH ARTICLE Open Access

A full-length transcriptome and gene expression analysis reveal genes and molecular elements expressed during seed development in *Gnetum luofuense*

Nan Deng[1,2†], Chen Hou[3,4†], Boxiang He[3,4], Fengfeng Ma[1,2], Qingan Song[1,2], Shengqing Shi[5], Caixia Liu[1*] and Yuxin Tian[1,2*]

Iso-Seq for genome annotation; RNA-Seq for DE analysis

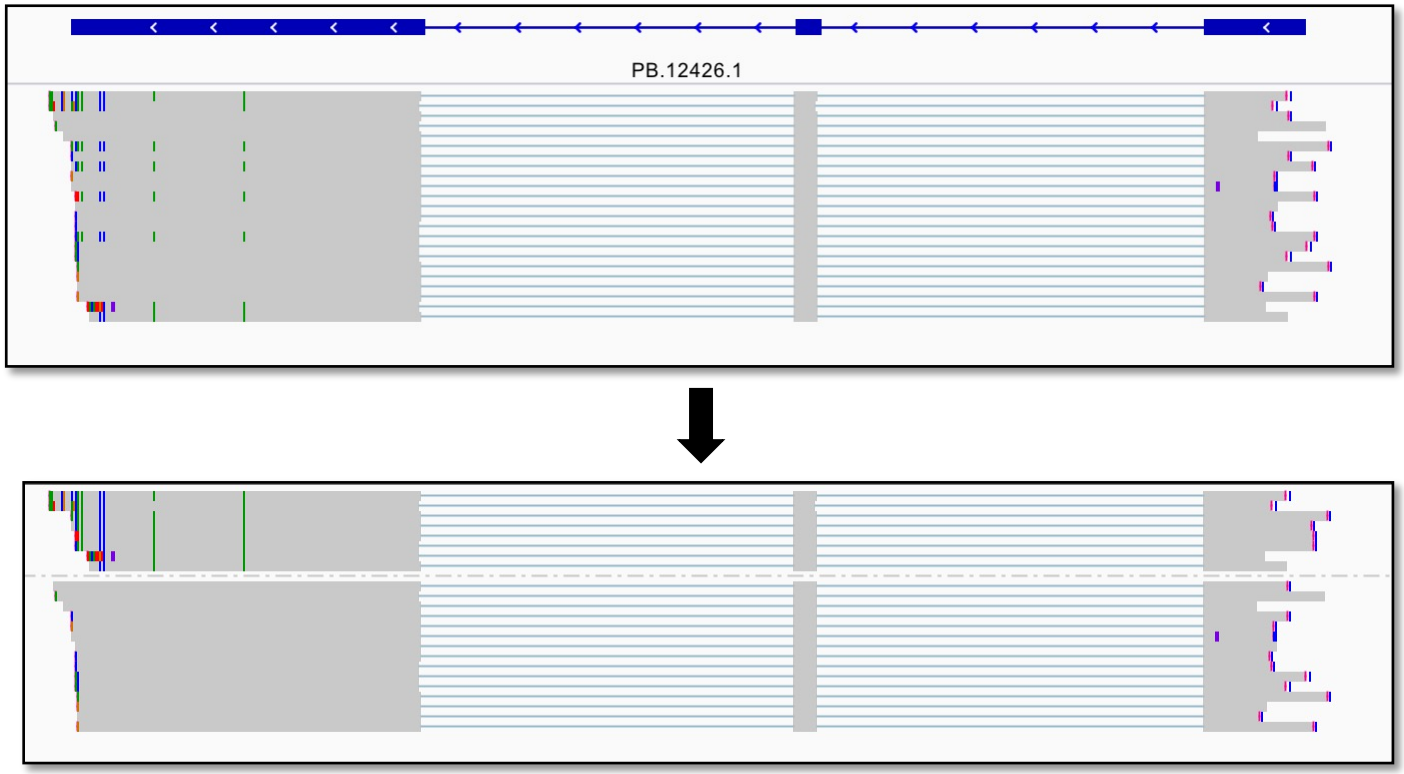
PACBIO®

Isoform-Level Phasing Using the Iso-Seq Method

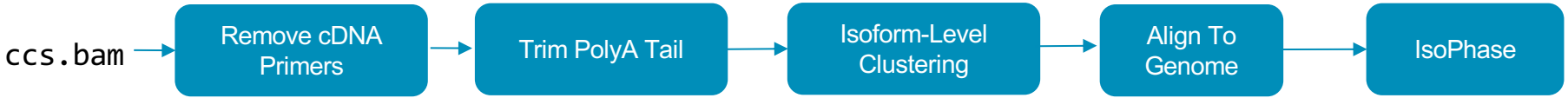
Application to diploid and polyploid species

ISOFORM-LEVEL PHASING USING ISO-SEQ READS

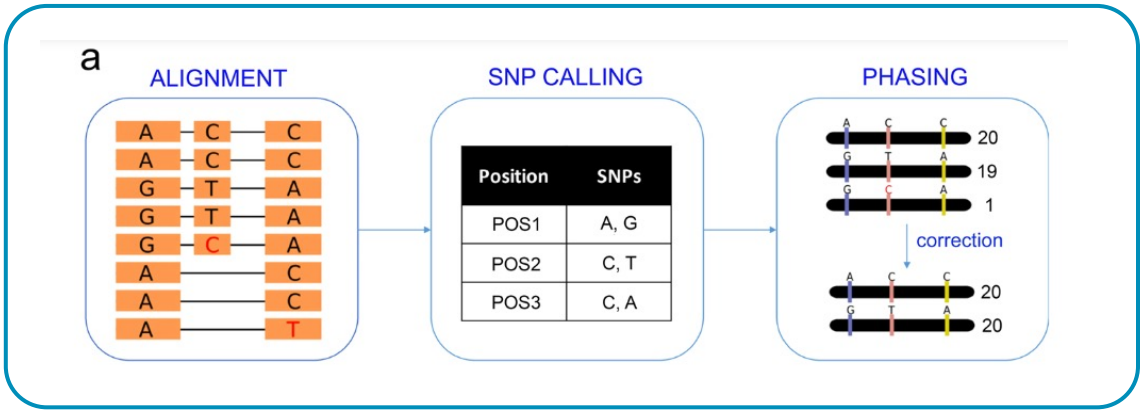
Individual reads represent single molecules that carry allele-specific information



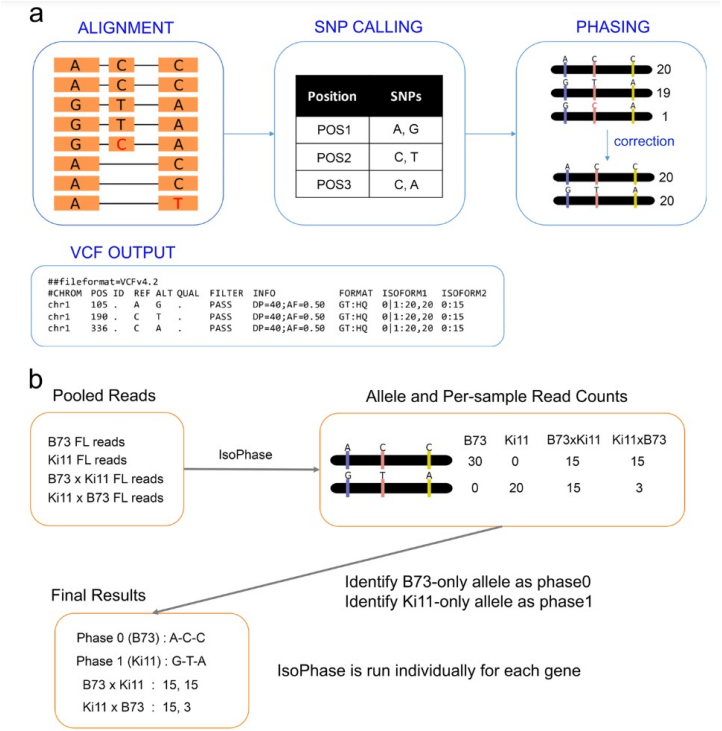
ISOPHASE V.0: ISOFORM-LEVEL PHASING

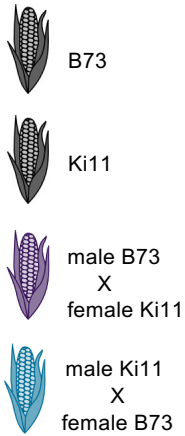


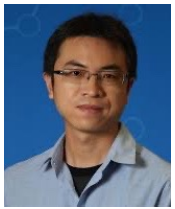
IsoPhase version 0



ISOPHASE V.0: ISOFORM-LEVEL PHASING







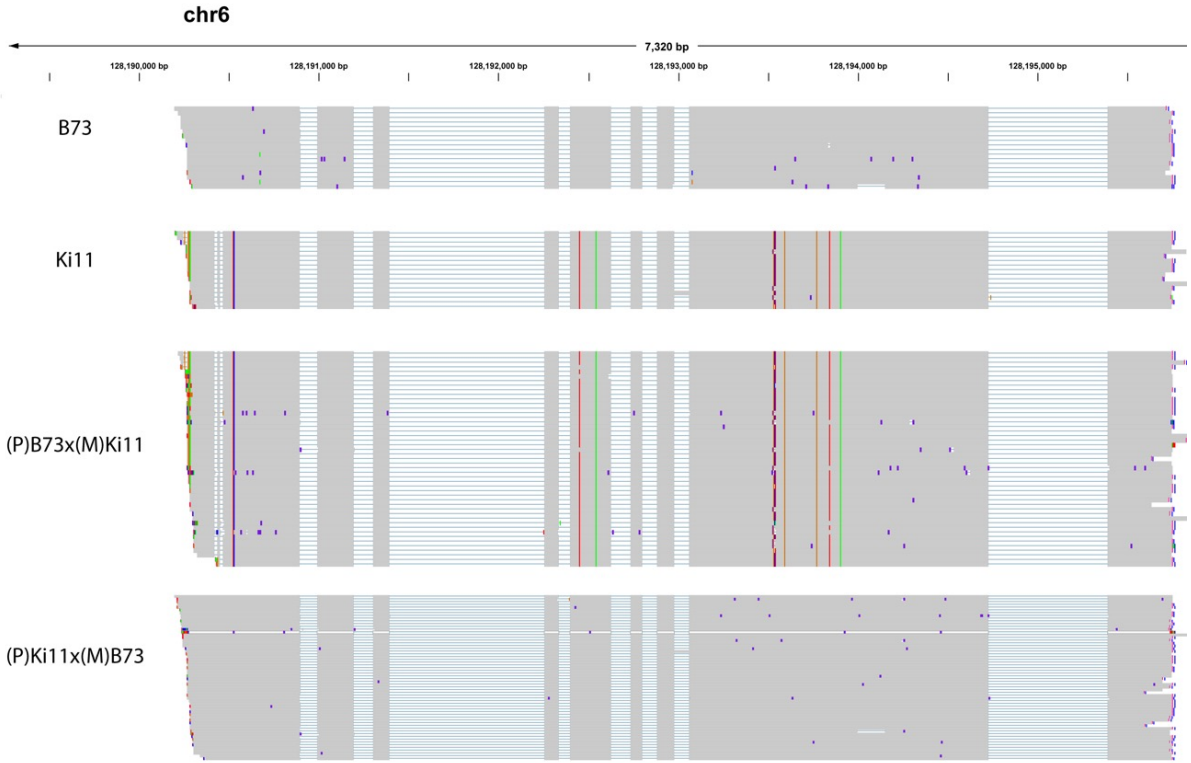
Bo Wang, CSHL

Doreen Ware, CSHL

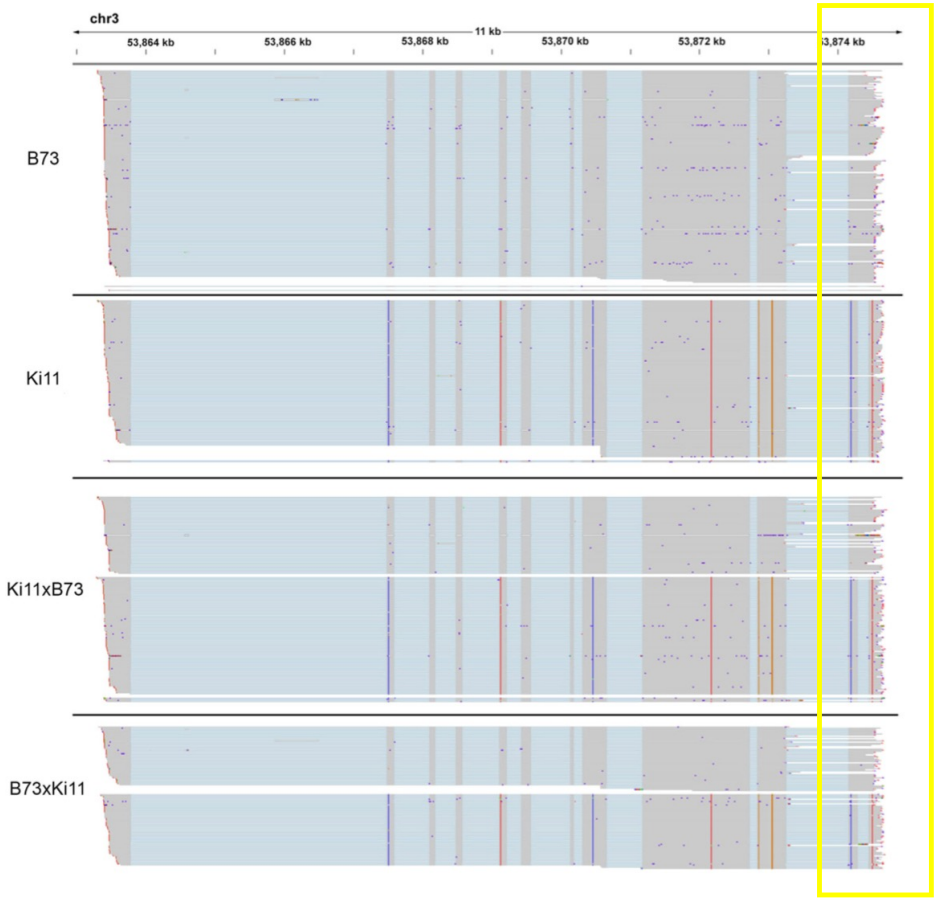
IsoPhase: https://github.com/Magdoll/cDNA_Cupcake/wiki/IsoPhase:-Haplotyping-using-Iso-Seq-data

Wang, B. et al. (2020) "Variant Phasing and Haplotypic Expression from Long-Read Sequencing in Maize." *Communications Biology*

MATERNAL IMPRINTING IN MAIZE



Wang, B. et al. (2020) "Variant Phasing and Haplotypic Expression from Long-Read Sequencing in Maize." *Communications Biology*



Parents express only one isoform

F1s express parent-specific isoforms

(3' UTR difference)

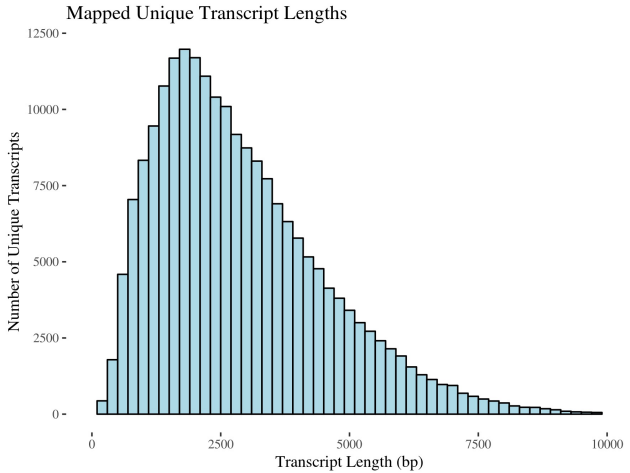
Wang, B. et al. (2020) "Variant Phasing and Haplotypic Expression from Long-Read Sequencing in Maize." *Communications Biology*

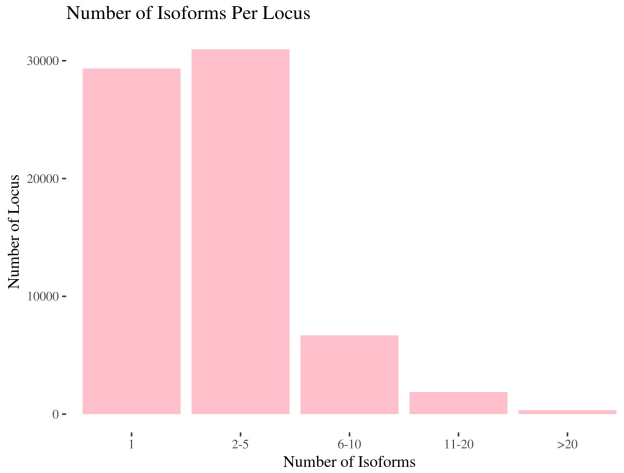
Sequencing the Coastal Redwood with the Iso-Seq Method

REDWOOD ISO-SEQ COLLECTION & ANALYSIS

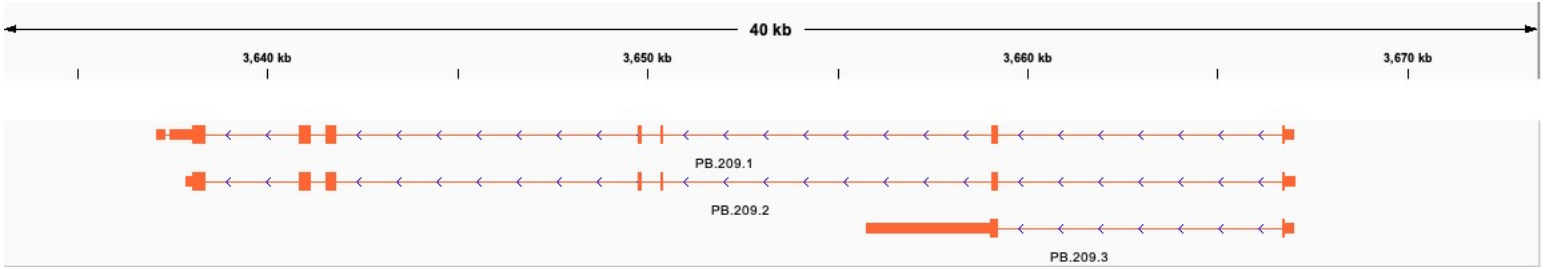
- Needles from same tree as PacBio genome
- Standard Iso-Seq analysis from SMRT Link
- Mapping to PacBio redwood v12 genome

	FL Reads	Unique Loci	Unique Transcripts
Redwood	5,379,490	69,198	205,792





ALTERNATIVE SPLICING DRIVES ORF CHANGES



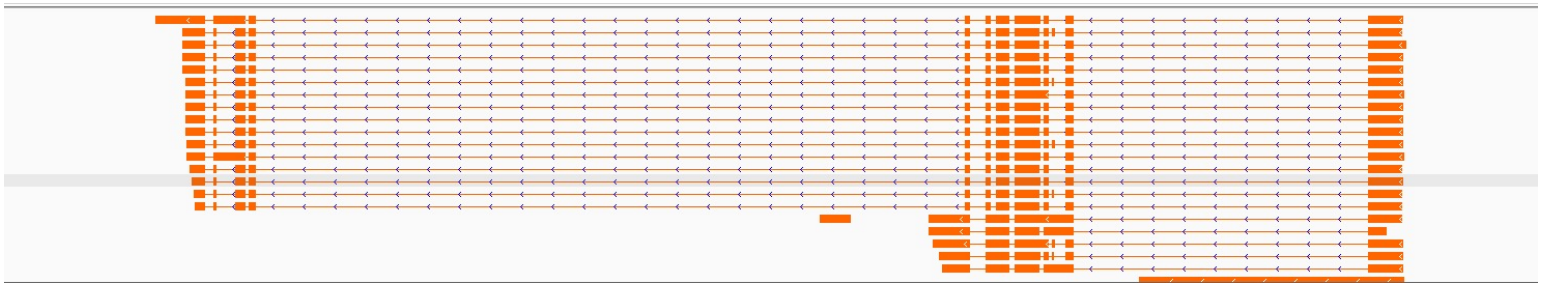

```
PB.209.1      MEYYGANSAMQRIDPASDWNLQAESSLEEAMRQMSMQSHDILQRGSGPYPERPGESDCAY60
PB.209.2      MEYYGANSAMQRIDPASDWNLQAESSLEEAMRQMSMQSHDILQRGSGPYPERPGESDCAY60
PB.209.3      MEYYGANSAMQRIDPASDWNLQAESSLEEAMRQMSMQSHDILQRGSGPYPERPGESDCAY60
              ************************************************************

PB.209.1      YMRNGVCGFGTNCRFNHPPNTNLGAPAARNRGEEYPERPGQPECQYFLKTGSCKFGATCK120
PB.209.2      YMRNGVCGFGTNCRFNHPPNTNLGAPAARNRGEEYPERPGQPECQYFLKTGSCKFGATCK120
PB.209.3      YMRNGVCGFGTNCRFNHPPNTNLVCFCLMLLVQID------LKSKWTL------------102
              ********************** . .     :         :.:: *

[omitted]


PB.209.1      DLRPEAAAGLSKEPISTSQAAPPSSTGEVSGSGALVTTSSNLAGSTFSGIDQK       473
PB.209.2      DLRPEAAAGLSKEPISTSQAAPPSSTGEVSGSGALVTTSSNLAGSTFSGIDQK       473
PB.209.3      --------------------------------------------------          102
```

BLASTP hit: zinc finger CCCH domain-containing protein 32-like

PHASING ISO-SEQ READS IN REDWOOD

PB.5798 ptg000465l:3375760-3416689

Iso-Seq transcripts



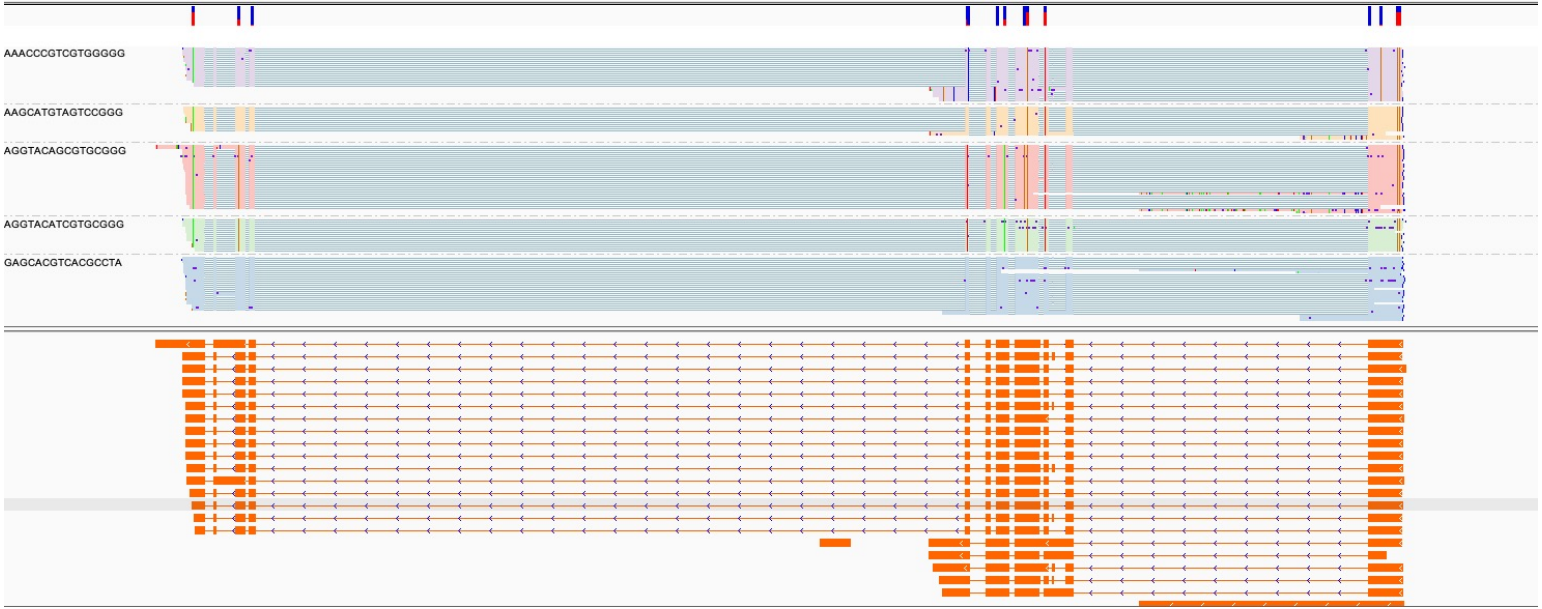
PHASING ISO-SEQ READS IN REDWOOD

16 SNPs identified; 5 alleles called

PB.5798 ptg000465l:3375760-3416689

Iso-Seq reads grouped by alleles

Iso-Seq transcripts

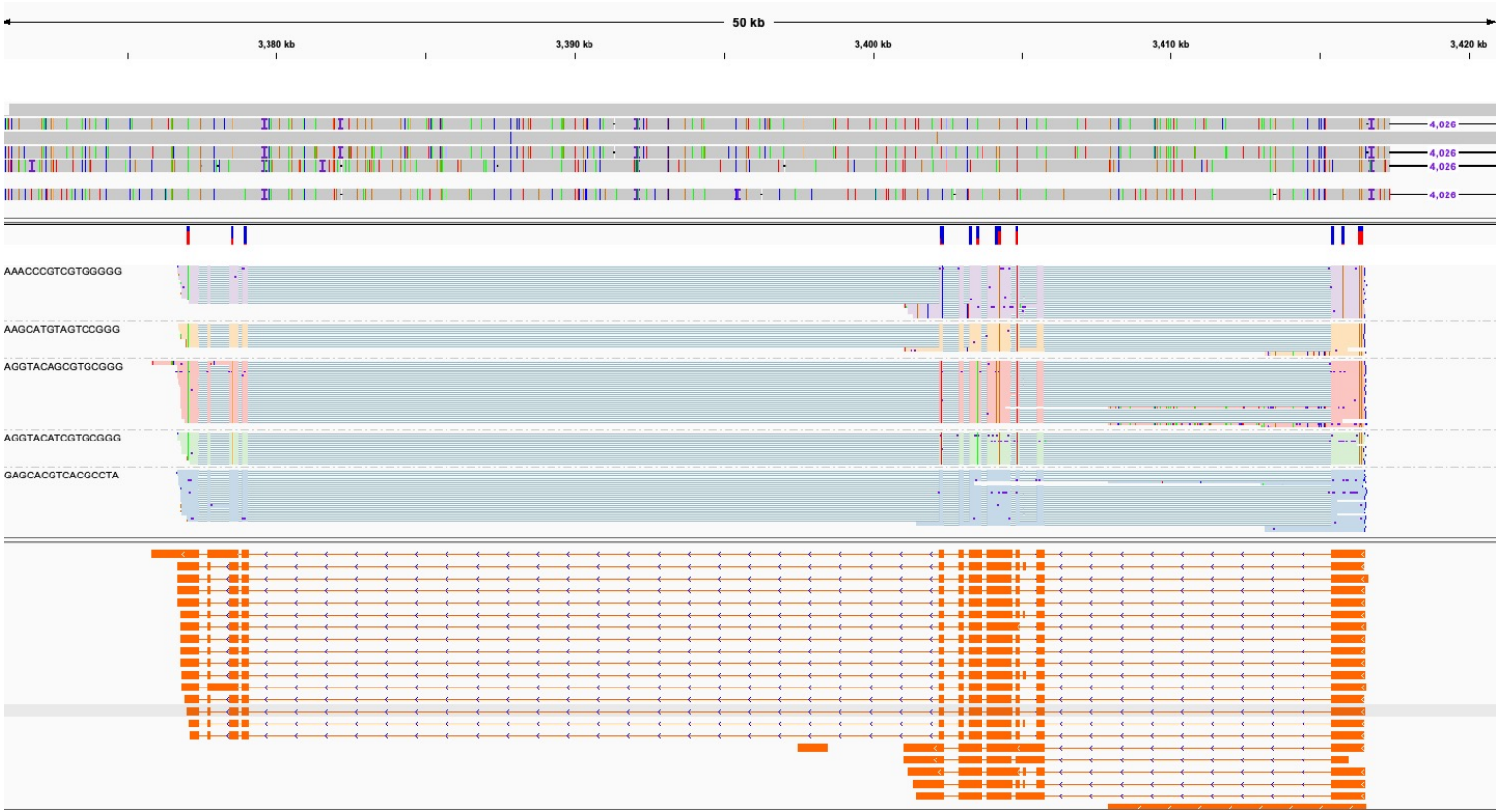


PB.5798 ptg000465l:3375760-3416689

Genome haplotigs

Iso-Seq reads grouped by alleles

Iso-Seq transcripts



https://downloads.pacbcloud.com/public/dataset/redwood2020/isoseq/

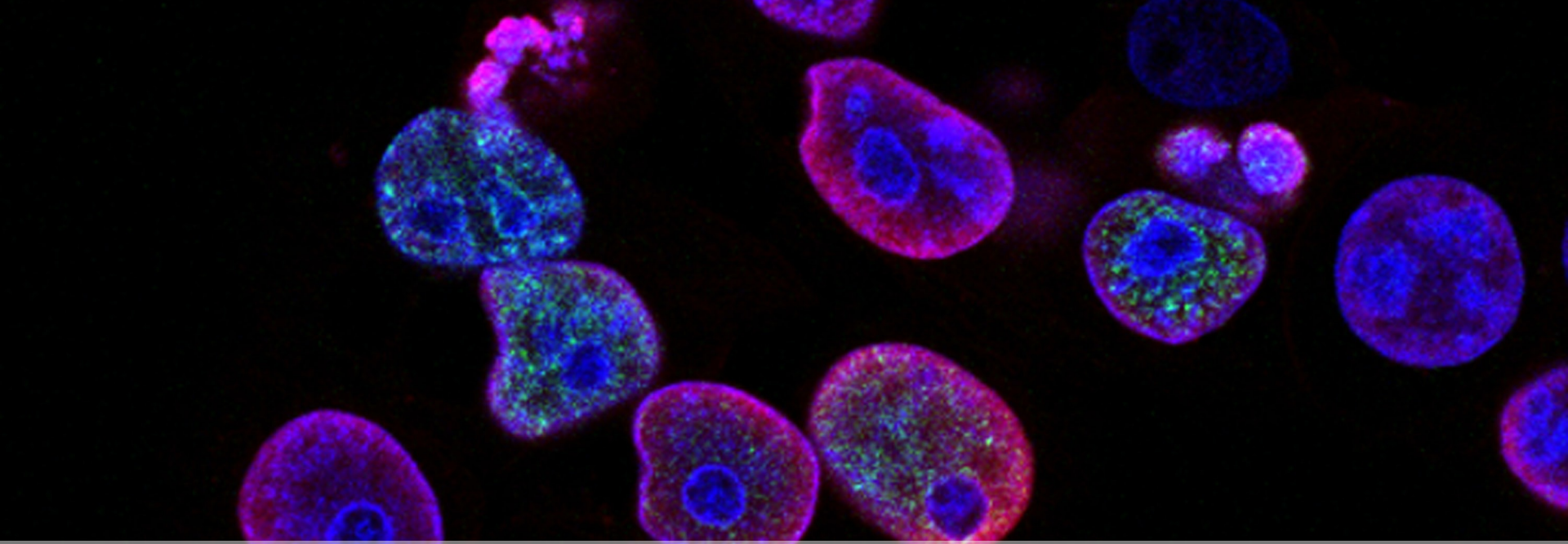
Index of /public/dataset/redwood2020/isoseq

Name	Last modified	Size	Description
Parent Directory		-	
Final-MappedTranscripts/	2021-01-06 12:42	-	
Final-UnmappedTranscripts/	2021-01-04 08:40	-	
Intermediate-FullLengthReads/	2020-12-28 13:17	-	
README.txt	2021-01-08 12:48	7.3K	

```
README  (Last Updated 01/08/2020)

Edited by: Elizabeth Tseng (etseng@pacb.com)


IMPORTANT: Please note that this release of Iso-Seq data maps to the
Hifiasm v12 version of the genome at
https://downloads.pacbcloud.com/public/dataset/redwood2020/hifiasm/v12/



********************
INTRODUCTION
********************

   This README file describes the contents in this directory.

   This dataset contains intermediate and processed files for an
Redwood Iso-Seq dataset. The library was sequenced on the Sequel II system
and processed using SMRTLink 10.1 followed by community tool analysis.

For more  information on Iso-Seq® methods[1], bioinformatics analysis,
see the PacBio Iso-Seq GitHub[2] and additional references below.


********************
SAMPLE
********************

Needles were collected from the same redwood tree as used for the genome sequencing,
flash frozen and stored at -80C.
```

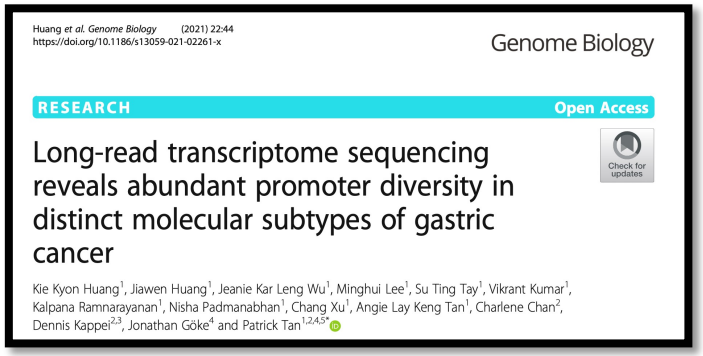
Redwood Iso-Seq Blog

https://medium.com/@Magdoll



Cancer & Rare Diseases

Iso-Seq Method in Gastric Cancer



Huang *et al. Genome Biology* (2021) 22:44
https://doi.org/10.1186/s13059-021-02261-x

Genome Biology

RESEARCH Open Access

Long-read transcriptome sequencing reveals abundant promoter diversity in distinct molecular subtypes of gastric cancer

Kie Kyon Huang[1], Jiawen Huang[1], Jeanie Kar Leng Wu[1], Minghui Lee[1], Su Ting Tay[1], Vikrant Kumar[1], Kalpana Ramnarayanan[1], Nisha Padmanabhan[1], Chang Xu[1], Angie Lay Keng Tan[1], Charlene Chan[2], Dennis Kappei[2,3], Jonathan Göke[4] and Patrick Tan[1,2,4,5*]

- Gastric cancer is the 3rd leading cause of cancer death
- Tumor morphology gives limited guidance
- Molecular methods (sequencing) can better help subtype GC for proper clinical treatment

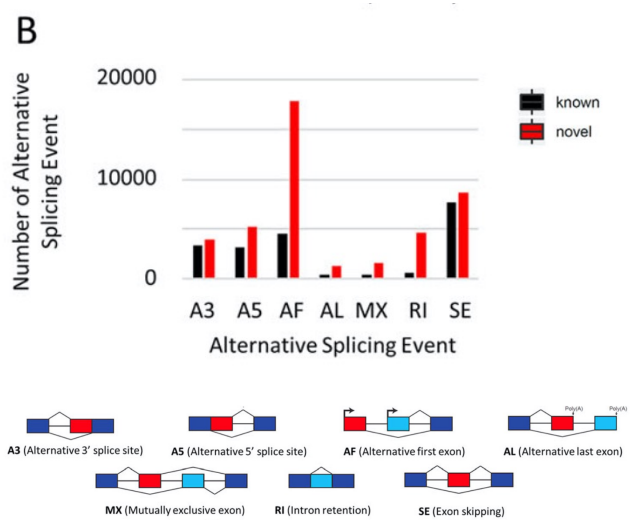
This is the first study that applies full-length transcript sequencing (Iso-Seq method) to gastric cancer for extensive characterization of alternative splicing and its potential for biomarker and drug discovery

Huang, Kie Kyon, Jiawen Huang, Jeanie Kar Leng Wu, Minghui Lee, Su Ting Tay, Vikrant Kumar, Kalpana Ramnarayanan, et al. 2021. “Long-Read Transcriptome Sequencing Reveals Abundant Promoter Diversity in Distinct Molecular Subtypes of Gastric Cancer.” *Genome Biology* 22 (1): 44.

Iso-Seq Method in Gastric Cancer

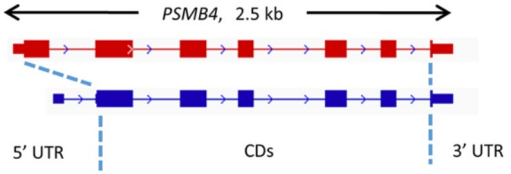
>60% Iso-Seq Transcripts Are Novel

Majority of novelty comes from use of an alternative first exon (AF)

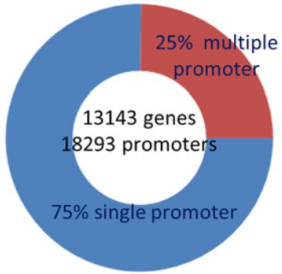


Alternative Promoters Change CDS

AFs can change the encoded protein



~25% genes have multiple promoters



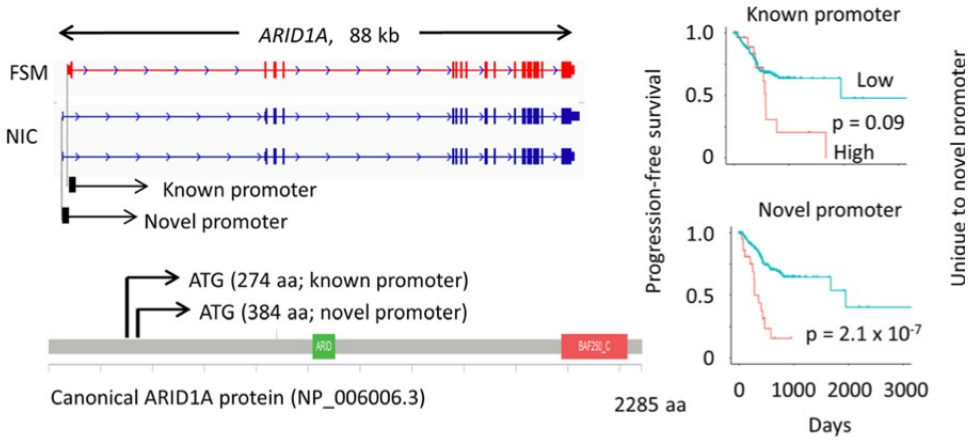
Huang, Kie Kyon, Jiawen Huang, Jeanie Kar Leng Wu, Minghui Lee, Su Ting Tay, Vikrant Kumar, Kalpana Ramnarayanan, et al. 2021. “Long-Read Transcriptome Sequencing Reveals Abundant Promoter Diversity in Distinct Molecular Subtypes of Gastric Cancer.” *Genome Biology* 22 (1): 44.

In Vitro Research: Linking novel promoters to potential clinical outcomes

Iso-Seq Method Identifies Novel Promoter in ARID1A Associated With Disease Prognosis

Two novel (NIC) transcripts use a novel promoter that truncates the first 384 aa; it is associated with poor survival outcome.

In contrast, the known (FSM) transcript uses a known promoter and is not significantly assoc. w poor survival.



Huang, Kie Kyon, Jiawen Huang, Jeanie Kar Leng Wu, Minghui Lee, Su Ting Tay, Vikrant Kumar, Kalpana Ramnarayanan, et al. 2021. “Long-Read Transcriptome Sequencing Reveals Abundant Promoter Diversity in Distinct Molecular Subtypes of Gastric Cancer.” *Genome Biology* 22 (1): 44.

Genetics of Rare Diseases

Circulation: Genomic and Precision Medicine

RESEARCH LETTER

Targeted Long-Read RNA Sequencing Demonstrates Transcriptional Diversity Driven by Splice-Site Variation in *MYBPC3*

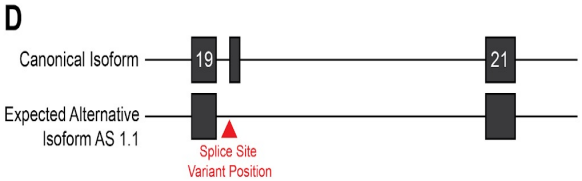
To date, clinical sequencing has focused on genomic DNA using targeted panels and exome sequencing. Sequencing of a large hypertrophic cardiomyopathy (HCM) cohort revealed that positive identification of a disease-associated variant was returned in only 32% of patients, with an additional 15% receiving inconclusive results.[1] When genome sequencing fails to reveal causative variants, the transcriptome may provide additional diagnostic clarity. A recent study examining patients with genetically undiagnosed muscle disorders found that RNA sequencing, when used as a complement to exome and whole genome sequencing, had an overall diagnosis rate of 35%.[2]

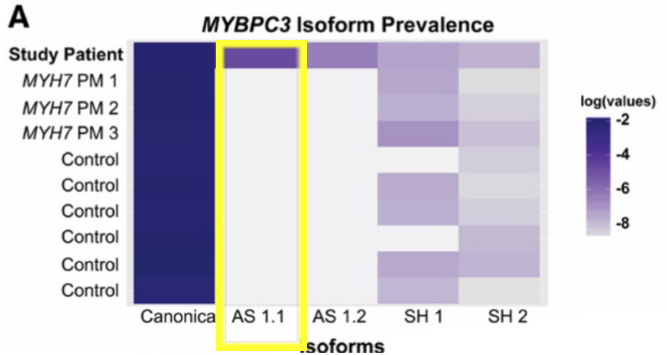
Alexandra Dainis, PhD
Elizabeth Tseng, PhD
Tyson A. Clark, PhD
Ting Hon, MS
Matthew Wheeler, MD, PhD
Euan Ashley, MB, ChB, DPhil

- 21y female patient with severe hypertrophic cardiomyopathy (HCM)
- HCM panel identified single base (c.1898-1G>A) mutation in *MYBPC3*
- Mutation expected to affect splicing between exon 19-20

Dainis et al., “Targeted Long-Read RNA Sequencing Demonstrates Transcriptional Diversity Driven by Splice-Site Variation in MYBPC3.” *Circulation. Genomic and Precision Medicine* (2019)

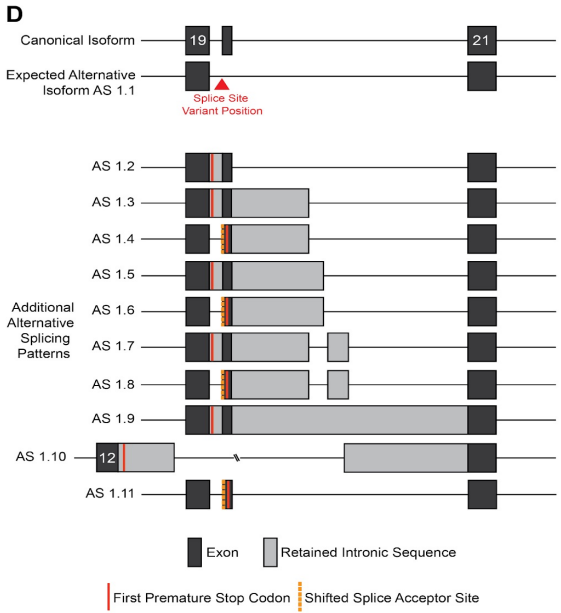
Aberrant splicing in MYBPC3 gene in a HCM patient

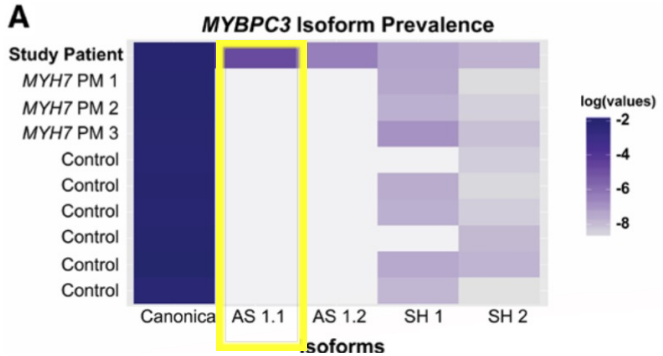




Dainis et al., “Targeted Long-Read RNA Sequencing Demonstrates Transcriptional Diversity Driven by Splice-Site Variation in MYBPC3.” *Circulation. Genomic and Precision Medicine* (2019)

Aberrant splicing in MYBPC3 gene in a HCM patient

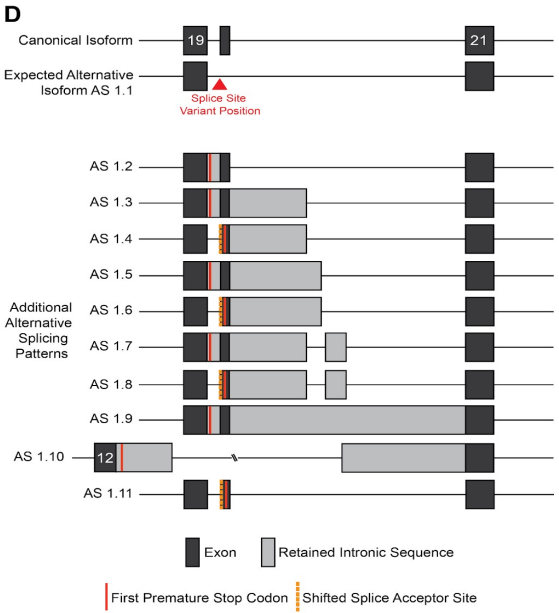


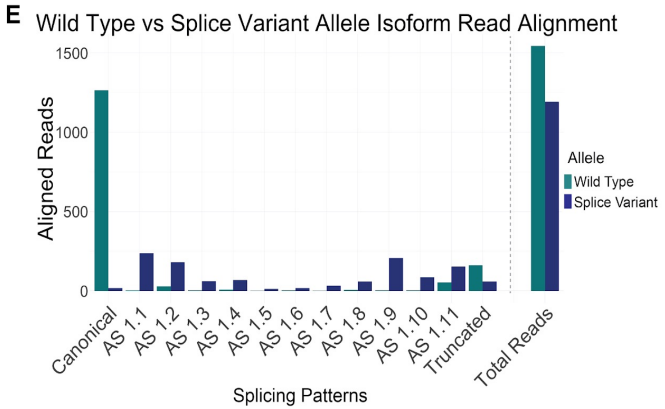


Dainis et al., “Targeted Long-Read RNA Sequencing Demonstrates Transcriptional Diversity Driven by Splice-Site Variation in MYBPC3.” *Circulation. Genomic and Precision Medicine* (2019)

Intronic SNP leads to aberrant splicing on the mutated allele

Iso-Seq identified patient-specific alternative splice variants (AS 1.1-1.11) linked to the c.1898-1G>A mutation not expressed on the WT allele





Dainis et al., "Targeted Long-Read RNA Sequencing Demonstrates Transcriptional Diversity Driven by Splice-Site Variation in MYBPC3." *Circulation. Genomic and Precision Medicine* (2019)

TOPICS TO BE DISCUSSED IN THE AFTERNOON

- Single-Cell Iso-Seq
- What to do without a reference genome
- Differential Analysis

Where To Learn More

pacb.com/applications/rna-sequencing/

tinyurl.com/isoseq-google

@PacBio @Magdoll

github.com/PacificBiosciences
github.com/Magdoll

PACBIO®

www.pacb.com