# The Iso-Seq Method for Human Diseases and Genome Annotation

Elizabeth Tseng/ June 2018

@Magdoll

PACBIO®

## SLIDES WILL BE POSTED ON TWITTER (@MAGDOLL) AND SPEAKERDECK.COM/MAGDOLL

Google Group:

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

GitHub Repository and Tutorials:

github.com/PacificBiosciences/IsoSeq_SA3nUP/

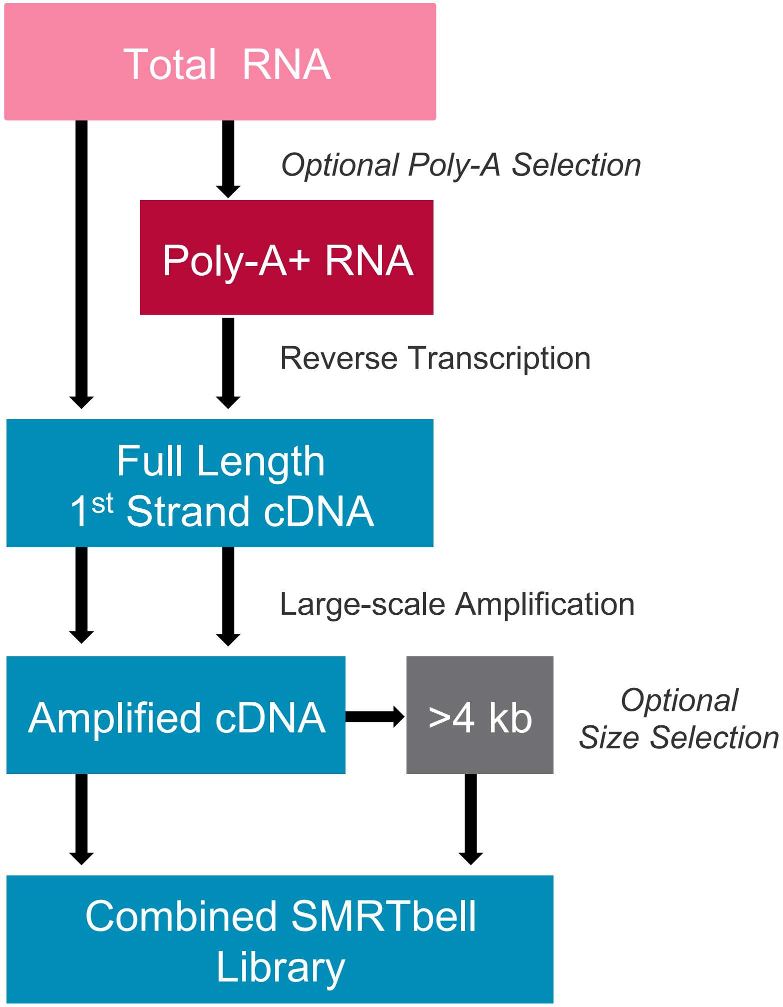
(**http://tinyurl.com/PBisoseq)**

# ISO-SEQ OVERVIEW

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

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

# SEQUEL ISO-SEQ LIBRARY PREPARATION



Total RNA

*Optional Poly-A Selection*

Poly-A+ RNA

Reverse Transcription

Full Length
1st Strand cDNA

Large-scale Amplification

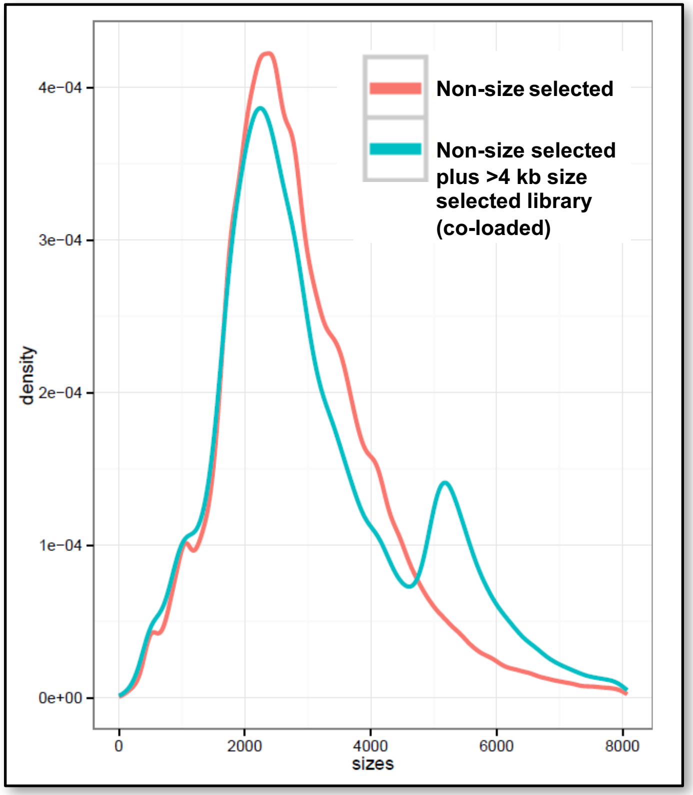
Amplified cDNA

>4 kb

*Optional*
*Size Selection*

Combined SMRTbell
Library

- Simplified library preparation
- Size selection optional



# OFFICIAL ISO-SEQ SOFTWARE SUPPORT

- SMRT Analysis (command line) / SMRT Link (GUI)
  - Latest Version: 5.1
  - Link : http://www.pacb.com/support/software-downloads/

**Main Features:**

- *de novo* (reference genome not required)
- no assembly required
- full-length (5’ to 3’)
- high accuracy ( >99%)

# Iso-Seq Publications Highlight

# ISO-SEQ PUBLICATIONS: WHOLE GENOME ANNOTATION

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 ~111k 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)

- Iso-Seq sequencing of maize and sorghum
- Comparative analysis of conserved and differentiated alternative splicing

# ISO-SEQ PUBLICATIONS: WHOLE GENOME ANNOTATION

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 ~60k high-quality transcripts (~29k genes)
- Identified > 20k potential lncRNAs

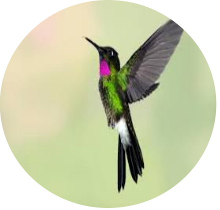
# ISO-SEQ PUBLICATIONS: WHOLE GENOME ANNOTATION

Cheng et al., **Long-read sequencing of the coffee bean transcriptome reveals the diversity of full-length transcripts.** *GigaScience* (2017)

- Obtained ~95k high-quality coffee bean transcripts
- Functional annotation using BLASTx, BLASTn, and BLAST2GO
- Identified new isoforms for caffeine-related genes

# ISO-SEQ PUBLICATIONS: WHOLE GENOME ANNOTATION

Workman et al., **Single molecule, full-length transcript sequencing provides insight into the extreme metabolism of ruby-throated hummingbird *Archilochus colubris*,** *GigaScience* (2018)



Jia et al., **SMRT sequencing of full-length transcriptome of flea beetle Agasicles hygrophila (Selman and Vogt).** *Sci. Rep.* (2018)

Wang et al., **A global survey of alternative splicing in allopolyploid cotton: landscape, complexity and regulation.** *New Phytol* (2017)

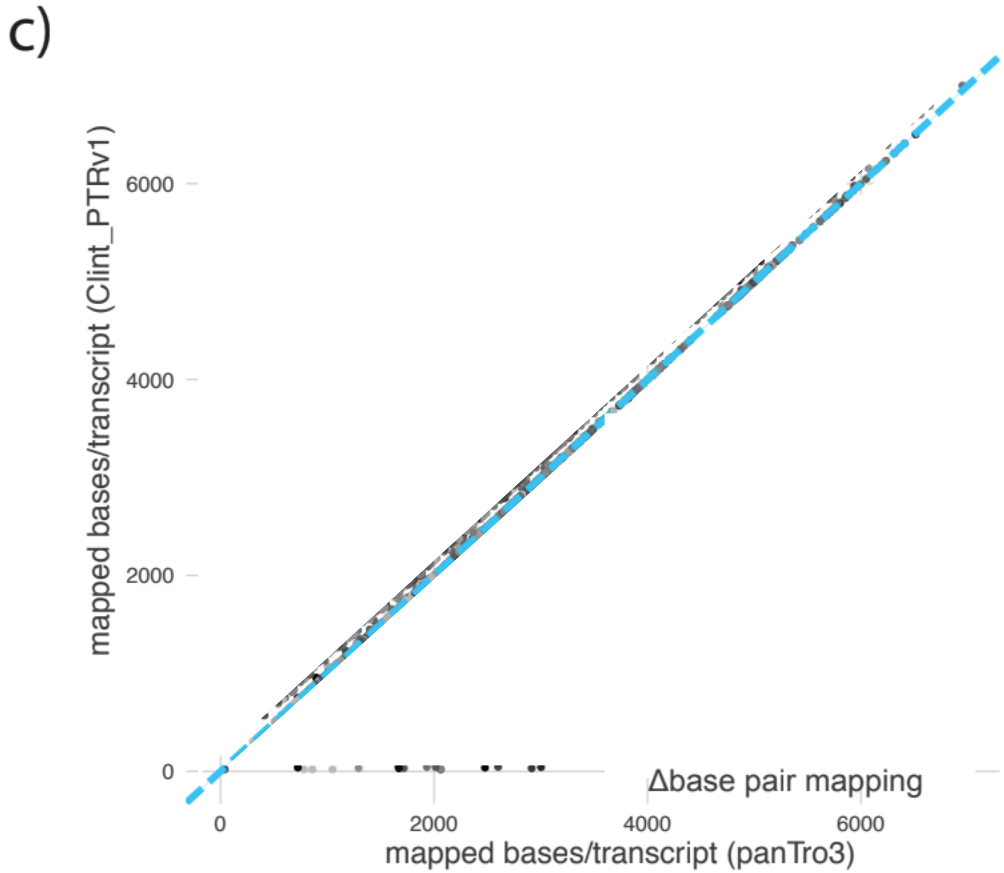
# COMPARATIVE GENOME + TRANSCRIPTOME SEQUENCING

Mexico City favors a scientist for mayor
New books to beat the summer heat
Carbon-based molecules on Mars
Science
$15
8 JUNE 2018
sciencemag.org
AAAS
GREAT APE
GENOMES
Deciphering humanity from our closest relatives

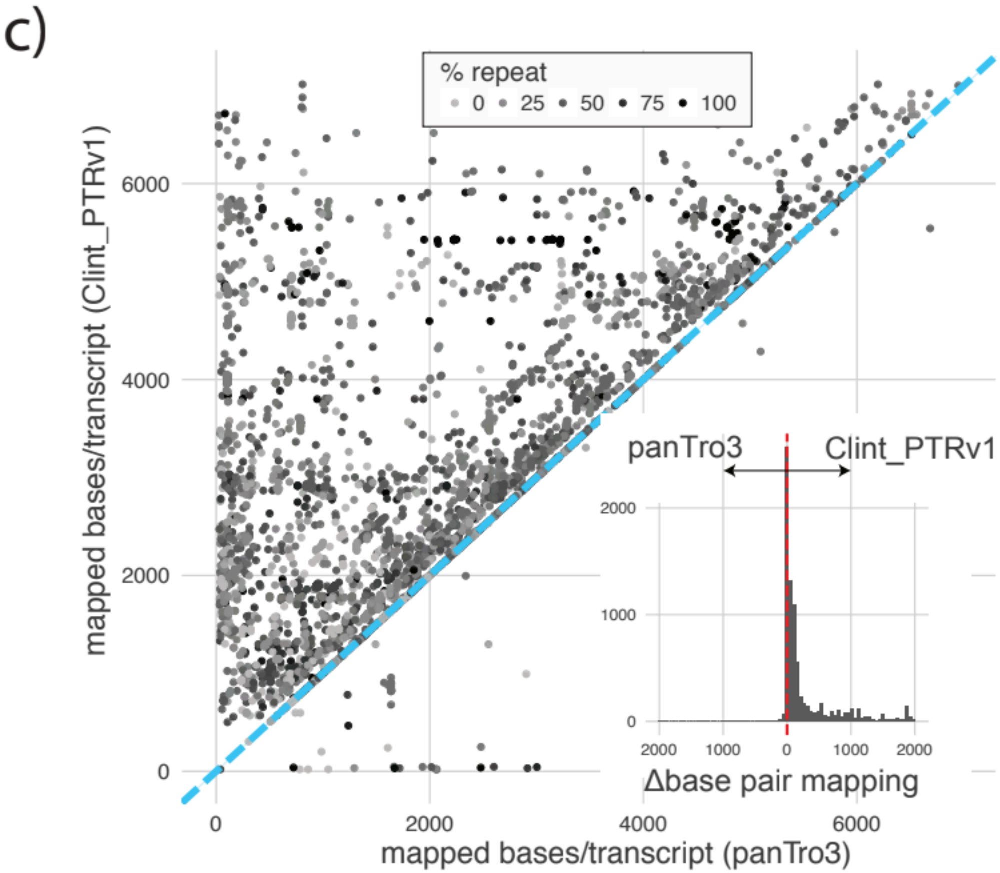
- Human, Chimp, and Orangutan
- *de novo* genome assembly using PacBio
- Iso-Seq + RNA-Seq for annotation

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

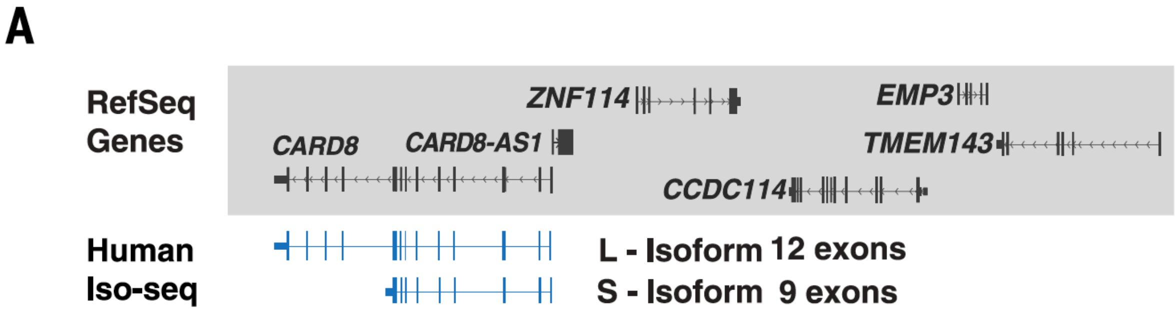
# CHIMP ASSEMBLY GREATLY IMPROVED



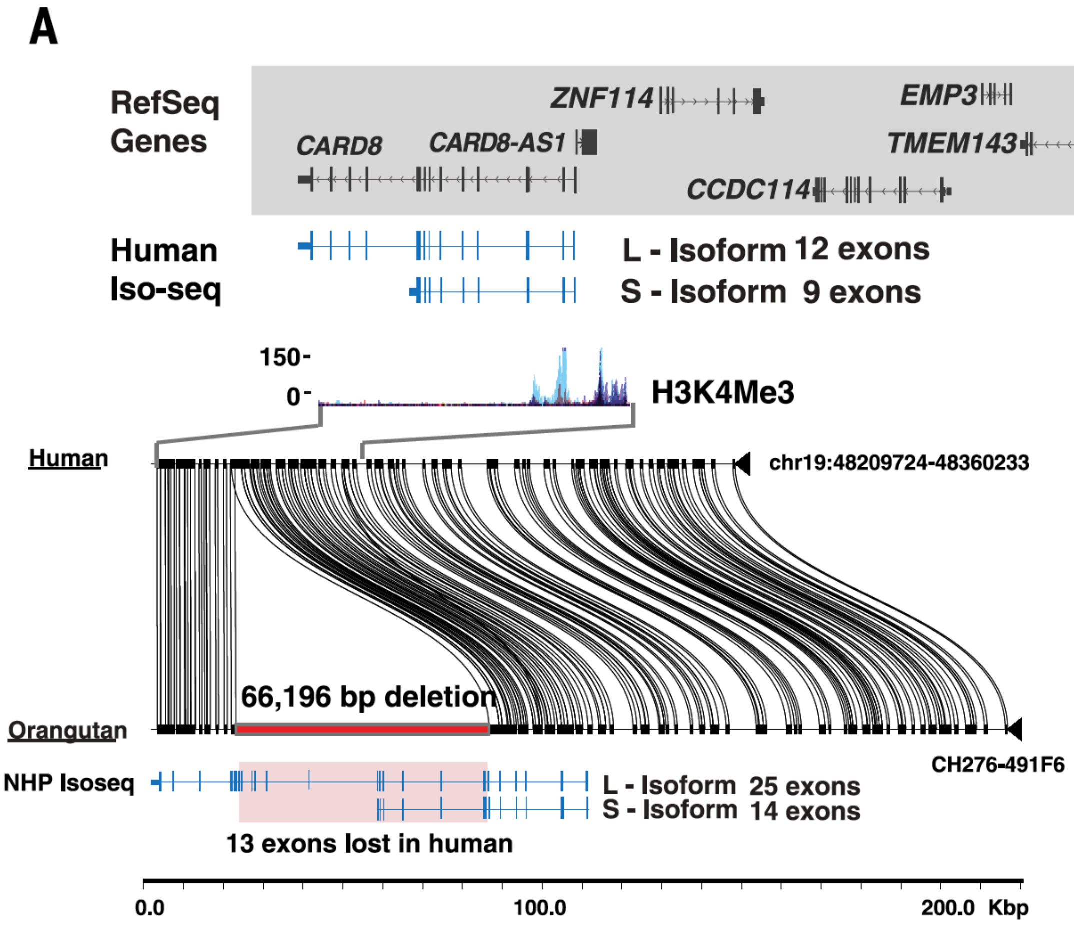
# CHIMP ASSEMBLY GREATLY IMPROVED



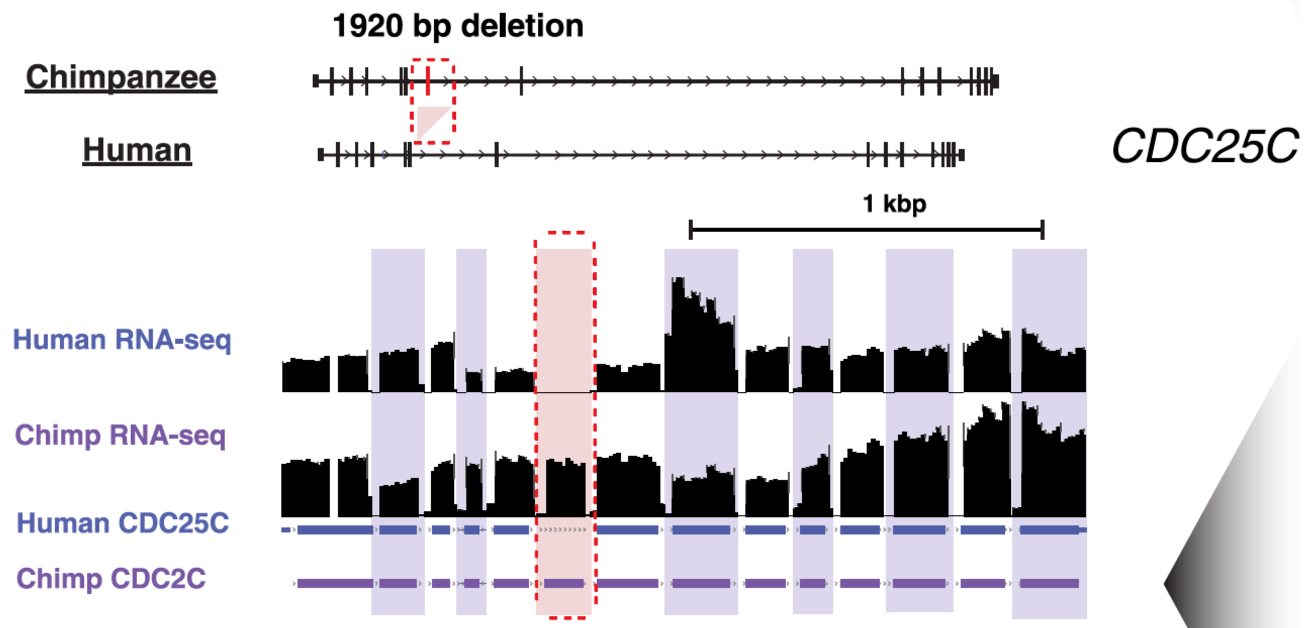
# HUMAN SPECIFIC DELETIONS DETECTED BY CROSS-SPECIES ISO-SEQ COMPARISON



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



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

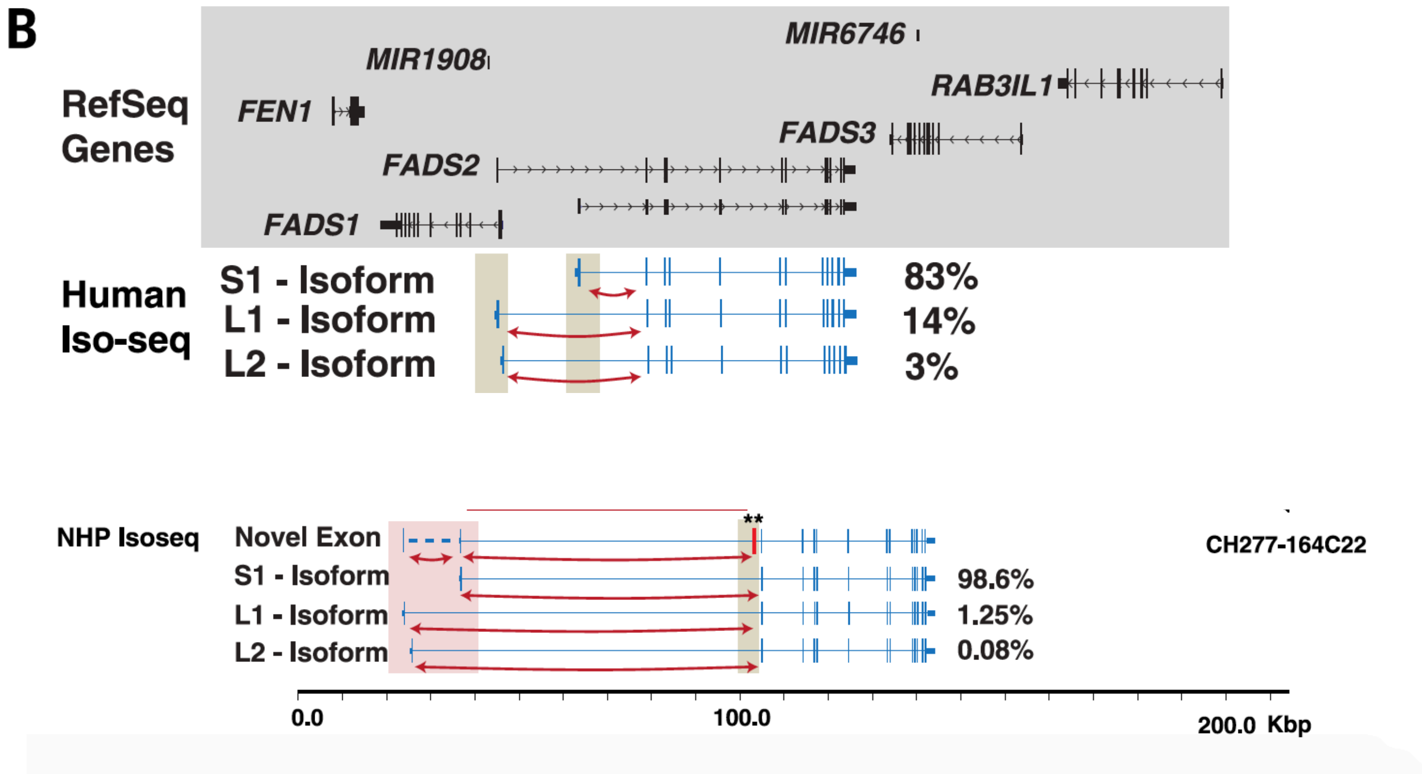


**Protein Alignment**

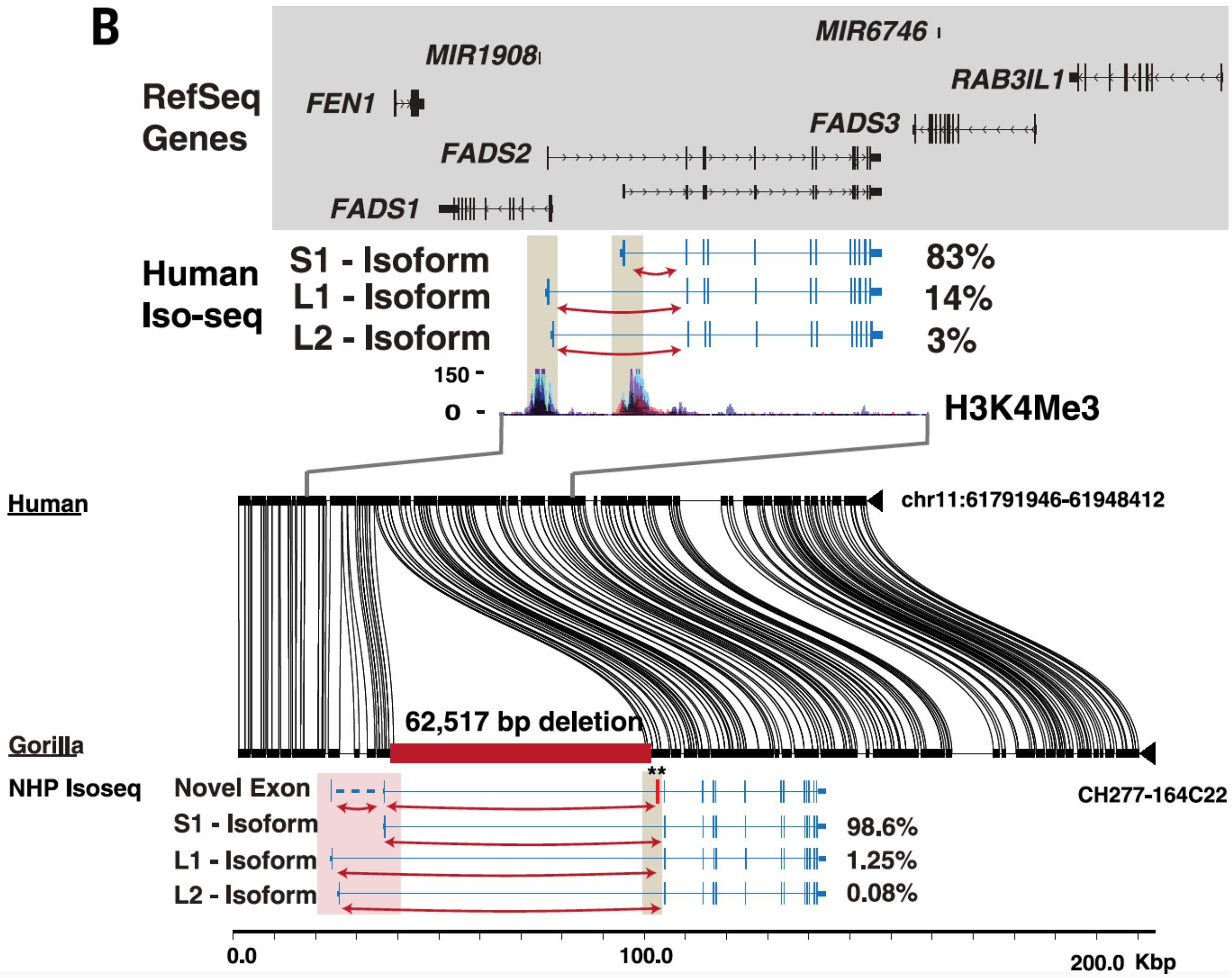
```
Green Monkey   KENENSTLKPLECQAPRNLRFQKRRGEPYVSPLSLLDNGNLVDSEMKYLG
Macaque        KENENGTLKPLECQAPRNLRFQKRPGEPYVSPLSLLDNGKLVDSEMKYLG
Gibbon         KENENGTLKPLECQAPRNLRFQKRPGEPYVSPLSLLDNGNLVDSEMKYLG
Orangutan      KENENSTLKPLEWQAPRNLRFQKRPGKPYVSPLSLLDNGNLVDSEMKYLG
Gorilla        KENENSTLKPLEWQAPRNLRFQKRPGKPYVSPLSLLDNGNLVDSEMKYLG
Chimpanzee     KENENSTLKPLEWQAPRNLRFQKRPGKPYVSPLSLLDNGNLVDSEIKYLG
Bonobo         KENENSTLKPLEWQAPRNLRFQKRPGKPYVSPLSLLDNGNLVDSEMKYLG
Human          KEN---------------------------------DNGNLVDSEMKYLG
```

**human-specific deletion - 33 AA**

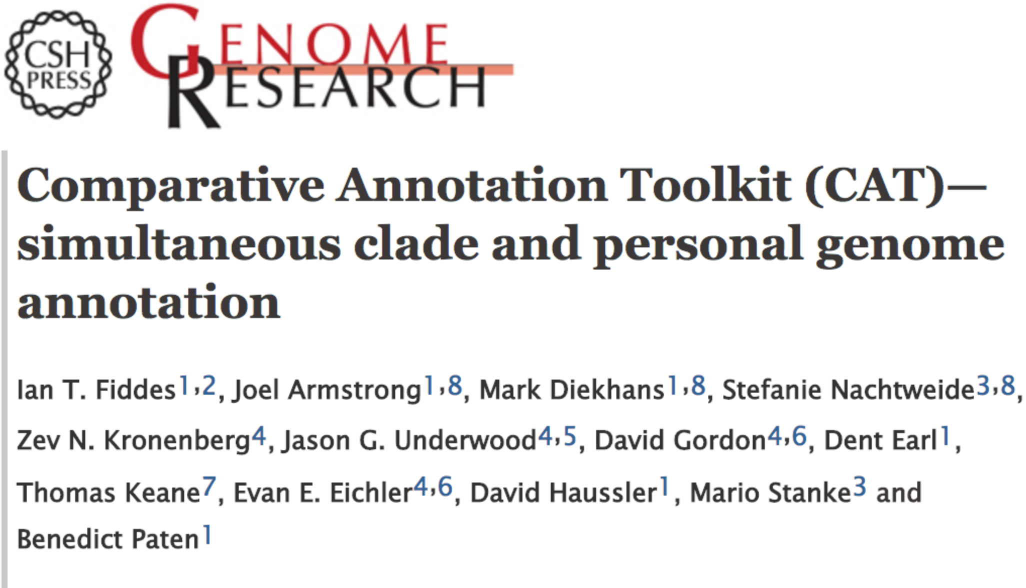
# 62KB DELETION IN FADS2 CHANGES EXONIC USAGE



# 62KB DELETION IN FADS2 CHANGES EXONIC USAGE



# CAT: COMPARATIVE ANNOTATION TOOLKIT



CSH PRESS GENOME RESEARCH

## Comparative Annotation Toolkit (CAT)—simultaneous clade and personal genome annotation

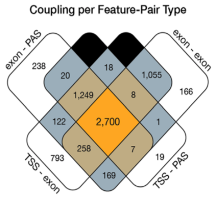
Ian T. Fiddes[1,2], Joel Armstrong[1,8], Mark Diekhans[1,8], Stefanie Nachtweide[3,8], Zev N. Kronenberg[4], Jason G. Underwood[4,5], David Gordon[4,6], Dent Earl[1], Thomas Keane[7], Evan E. Eichler[4,6], David Haussler[1], Mario Stanke[3] and Benedict Paten[1]

*…[CAT] provides a flexible way to **simultaneously annotate entire clades and identify orthology relationships**…resulting discovery of novel genes, isoforms, and structural variants….*

# ISO-SEQ PUBLICATIONS: HUMAN GENES AND DISEASES

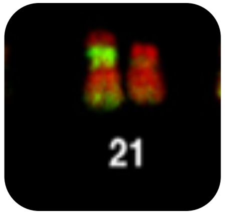
Treutlein et al., **Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing.** *Proc Natl Acad Sci* (2014)

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)

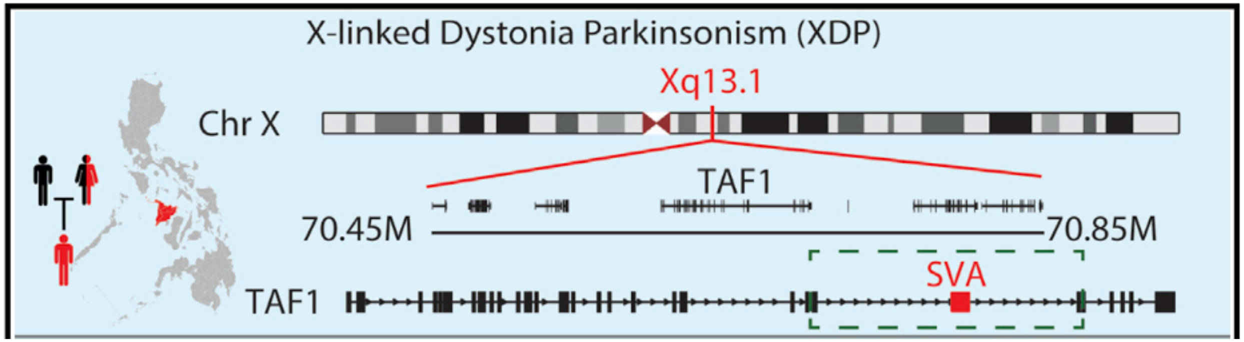
- AR-V7 is a known variant that prohibits successful therapy in castration-resistant prostate cancer
- RNA-seq data identified multiple AR variants, but unable to fully characterize

- 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
- Clinical data showed high AR-V9 expression predictive of therapy resistance

# ISO-SEQ HELPS SOLVE A RARE DISEASE

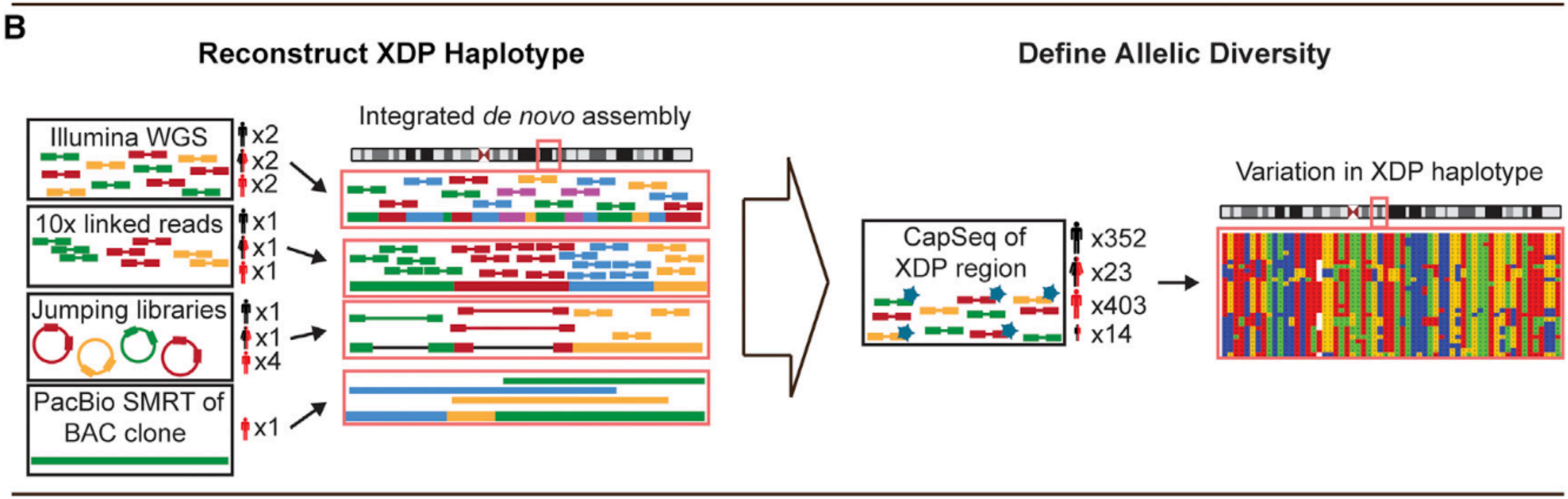
Aneichyk, T. *et al.* **Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly.** *Cell* (2018)

- X-linked Distonia-Parkinsonism (XDP) is a Mendelian neurodegenerative disease
- Endemic to Philippines Panay (6 in 100,000)
- Recent studies located causal variant in the TAF1 region on chrX
  - 5  single nucleotide variant (SNV)
  - 1 48-bp deletion
  - 1 2.6 kb SINE-VNTR-Alu (SVA) retrotransposon insertion



Aneichyk, T. *et al.* Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. *Cell* **172,** 897–902.e21 (2018).

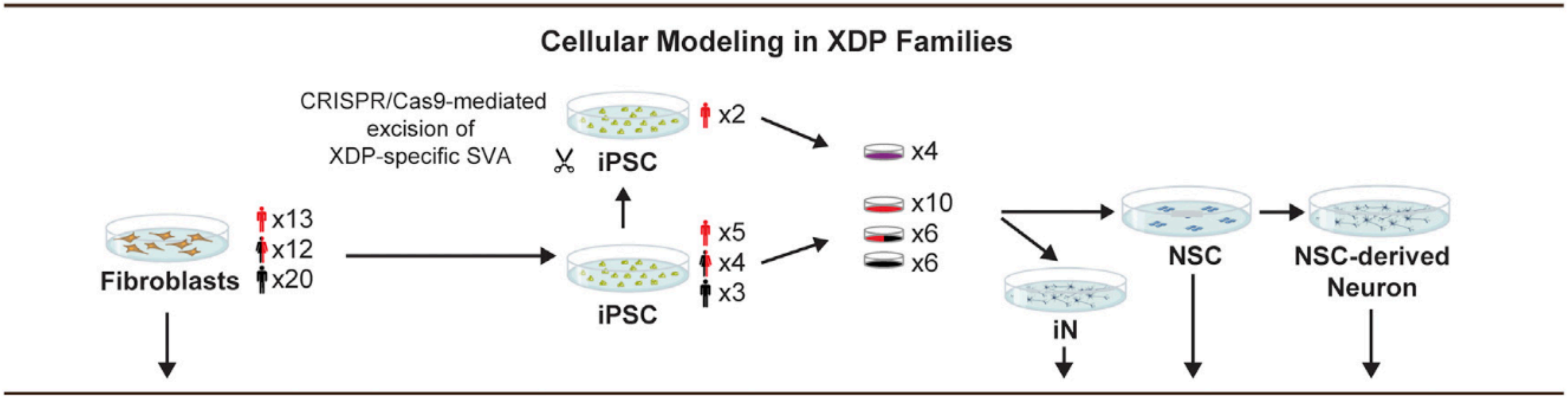
# ISO-SEQ HELPS SOLVE A RARE DISEASE



- First, de novo WGS to explore causal variants
  - Illumina + 10X
  - Long-insert jumping library (liWGS)
  - PacBio BAC cloning
  - Targeted capture of XDP region (CapSeq)
- Identified 47 additional new variants. Narrowed causal region down to TAF1 gene.

Aneichyk, T. *et al.* Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. *Cell* **172,** 897–902.e21 (2018).

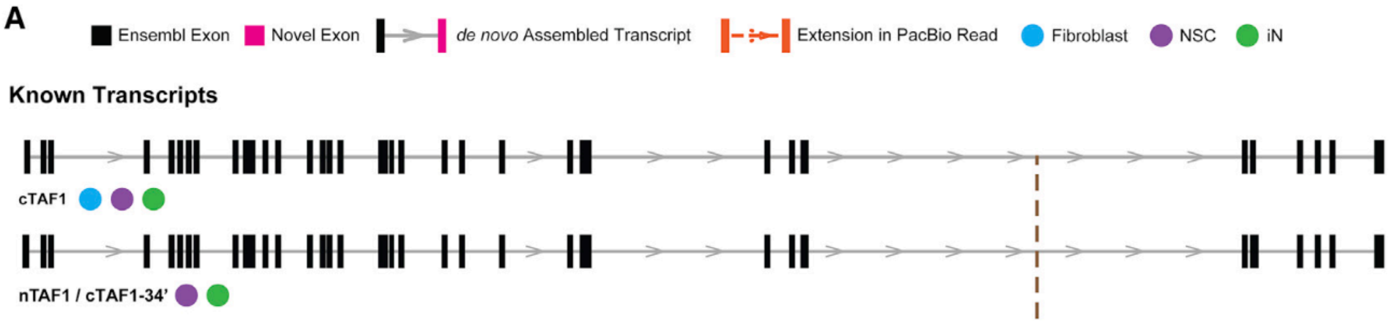
# ISO-SEQ HELPS SOLVE A RARE DISEASE


Cellular Modeling in XDP Families
CRISPR/Cas9-mediated excision of XDP-specific SVA
iPSC
x2
x4
x10
x6
x6
x13
x12
x20
Fibroblasts
iPSC
x5
x4
x3
NSC
NSC-derived Neuron
iN

- Transcriptome sequencing on XDP and control cell lines
  - Strand-specific RNA-seq
  - mRNA targeted capture (Illumina)
  - mRNA targeted capture (PacBio Iso-Seq)

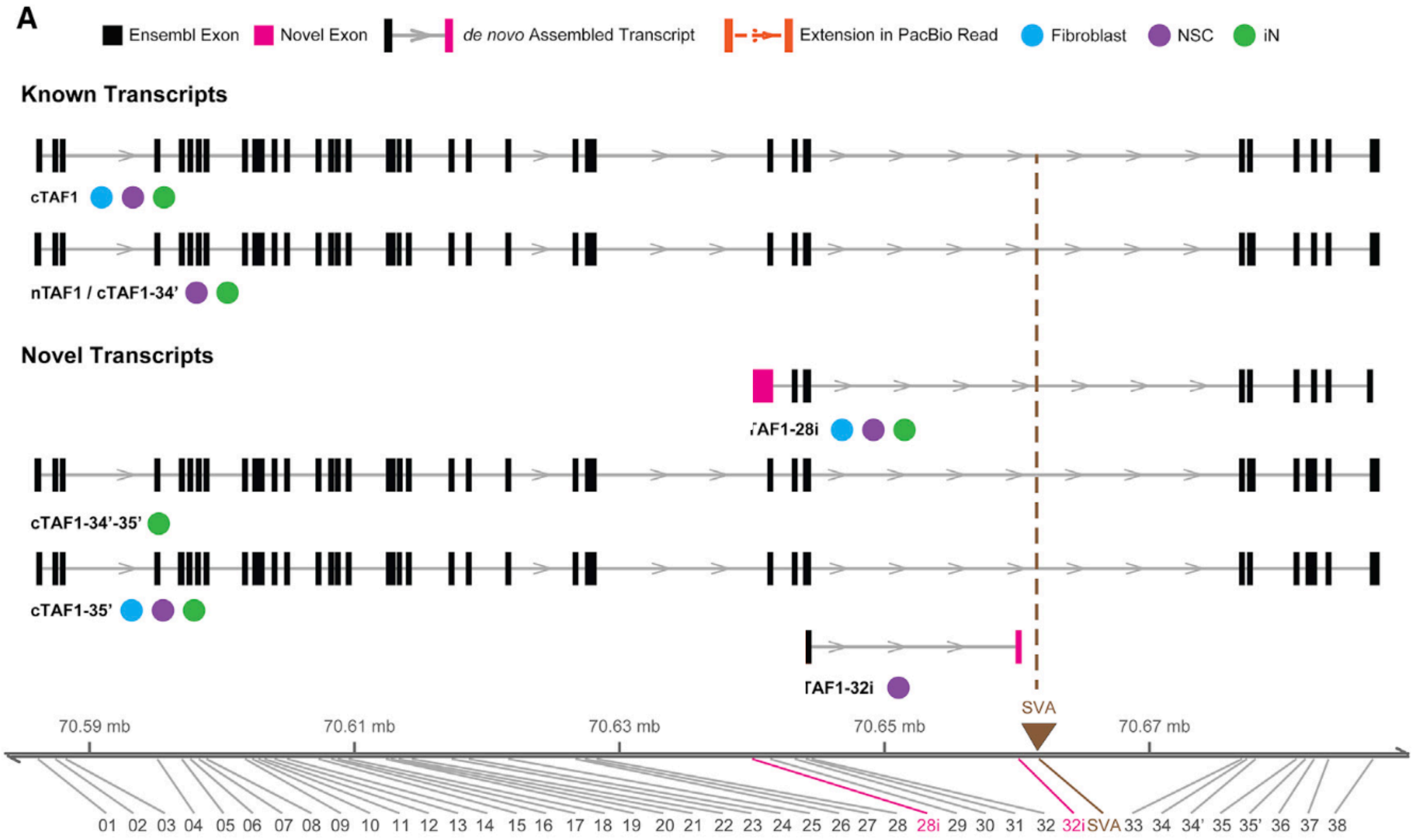
Aneichyk, T. *et al.* Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. *Cell* **172,** 897–902.e21 (2018).

# ISO-SEQ DATA EXTENDS 5’ END OF NOVEL ISOFORMS



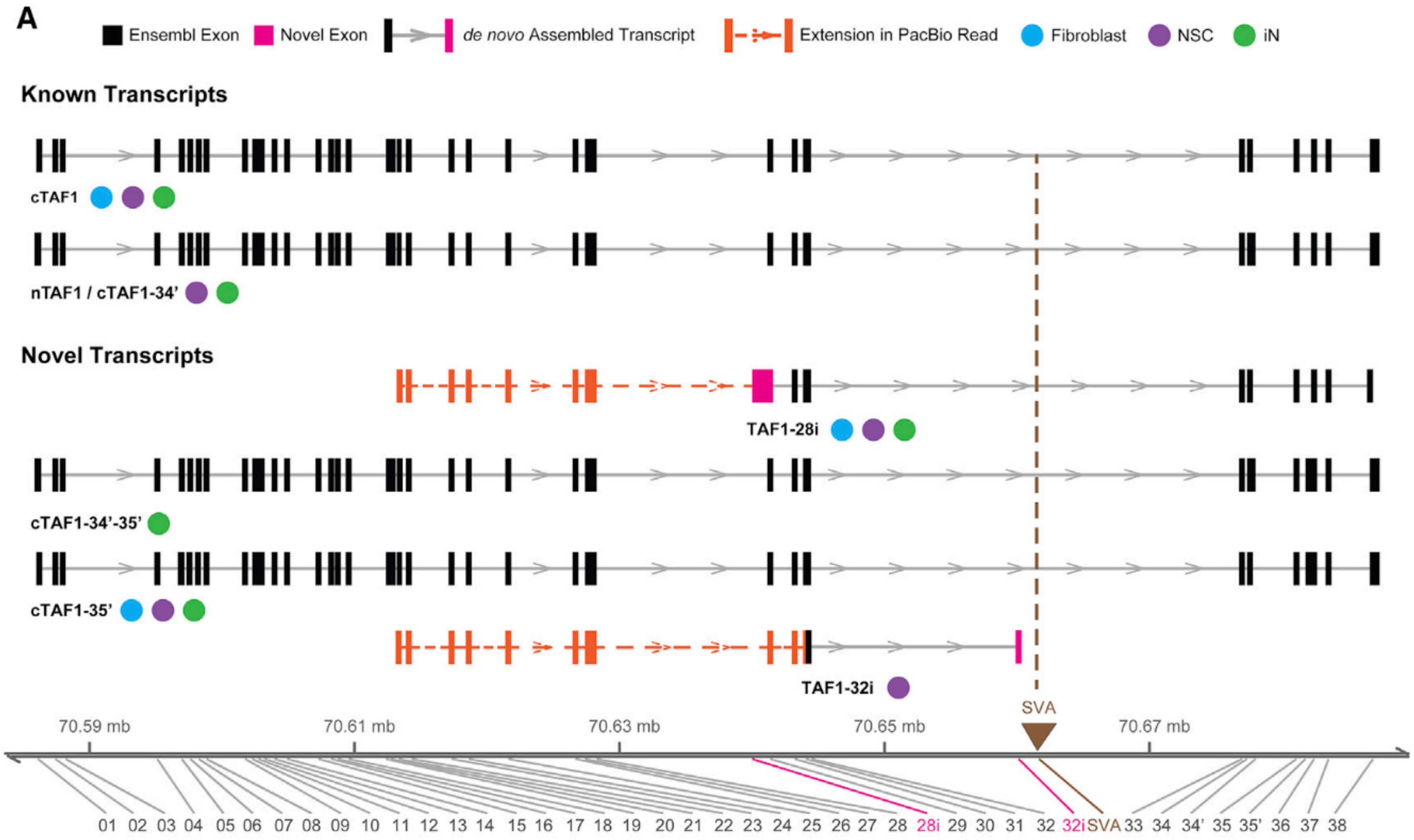
Aneichyk, T. *et al.* Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. *Cell* **172,** 897–902.e21 (2018).

# ISO-SEQ DATA EXTENDS 5’ END OF NOVEL ISOFORMS



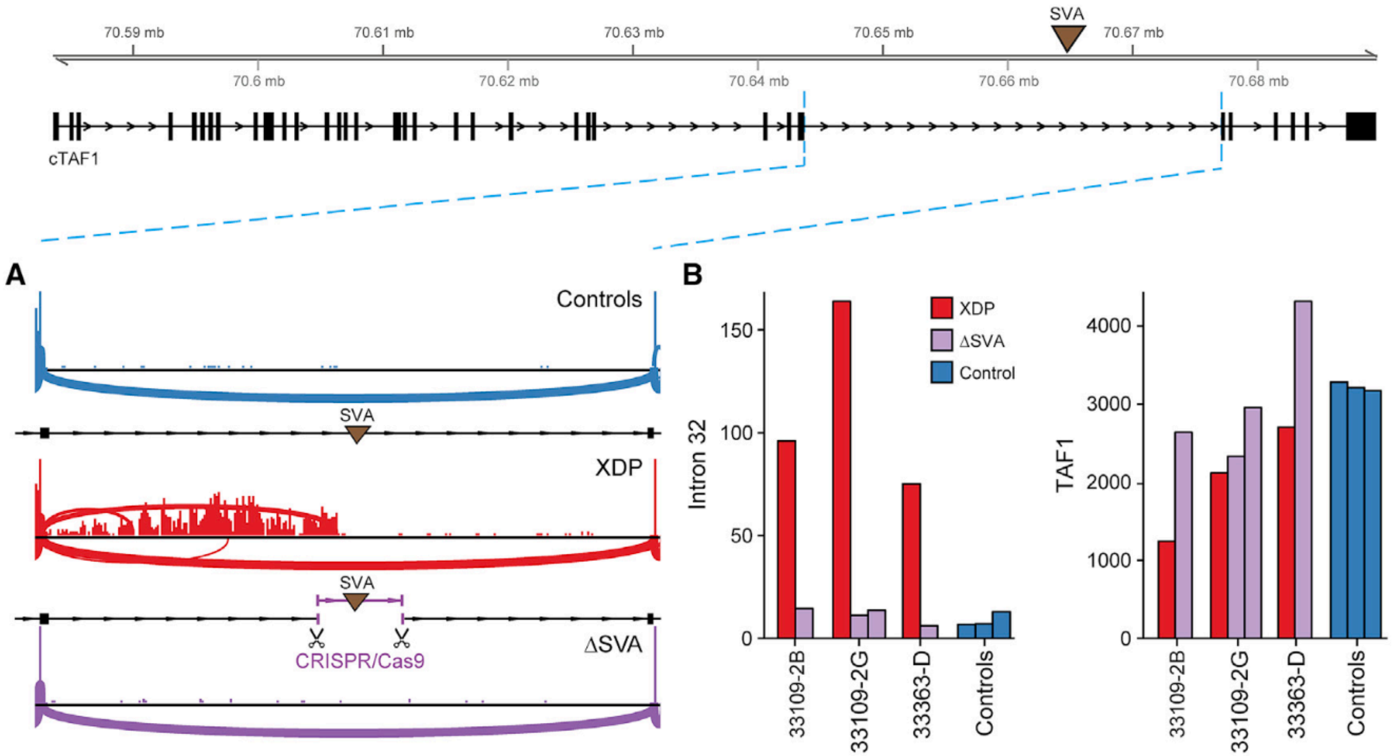
Aneichyk, T. *et al.* Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. *Cell* **172,** 897–902.e21 (2018).

# ISO-SEQ DATA EXTENDS 5’ END OF NOVEL ISOFORMS



Aneichyk, T. *et al.* Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. *Cell* **172,** 897–902.e21 (2018).

# CRISPR/CAS9 CONFIRMED SVA LINKED TO INTRON RETENTION



Aneichyk, T. *et al.* Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. *Cell* **172,** 897–902.e21 (2018).

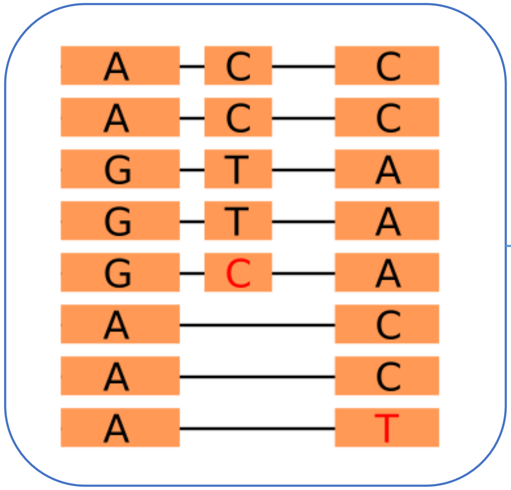
PACBIO®

# Iso-Phase

Using Iso-Seq data to phase isoforms

# ISOPHASE: ISOFORM PHASING USING ISO-SEQ DATA

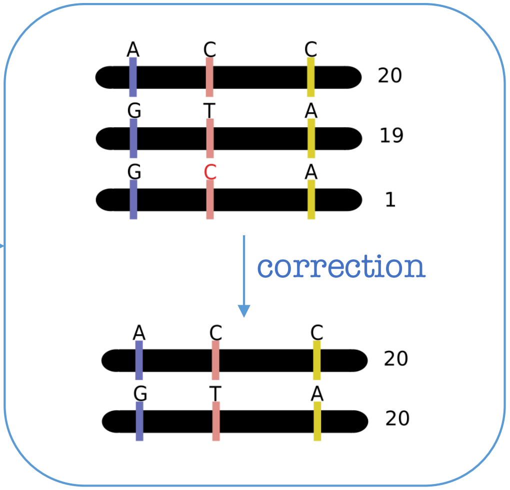
ALIGNMENT



SNP CALLING

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

PHASING



Can take optional RNA-seq input for SNP calling

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

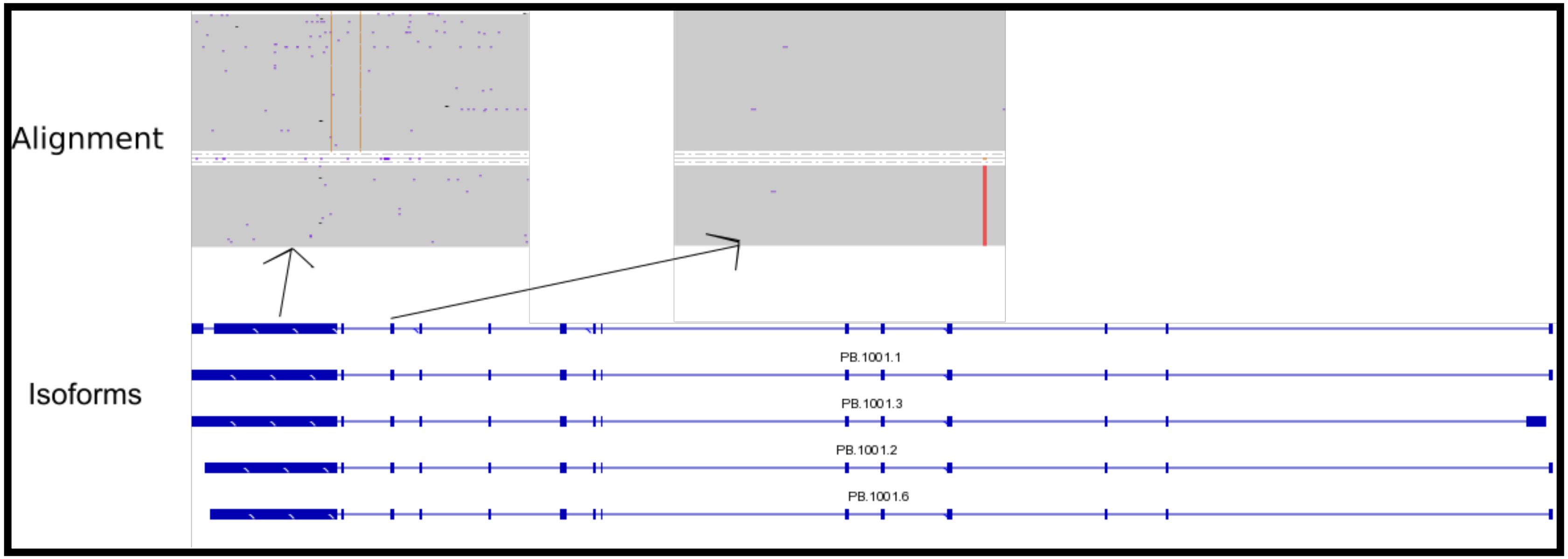
**Genome Assembly**

- Angus (sire) x Brahman (dam) F1 cattle
- PacBio, assembled with Falcon
- ~90% of genome phased using Unzip

**Iso-Seq Transcriptome Data**

- 30,137 final isoforms (12,101 genes)
- Selected for phasing: 1758 genes with ≥ 40 full-length CCS read coverage

# VPS36 ISOFORMS CALLED SNPS NOT PHASED IN GENOME

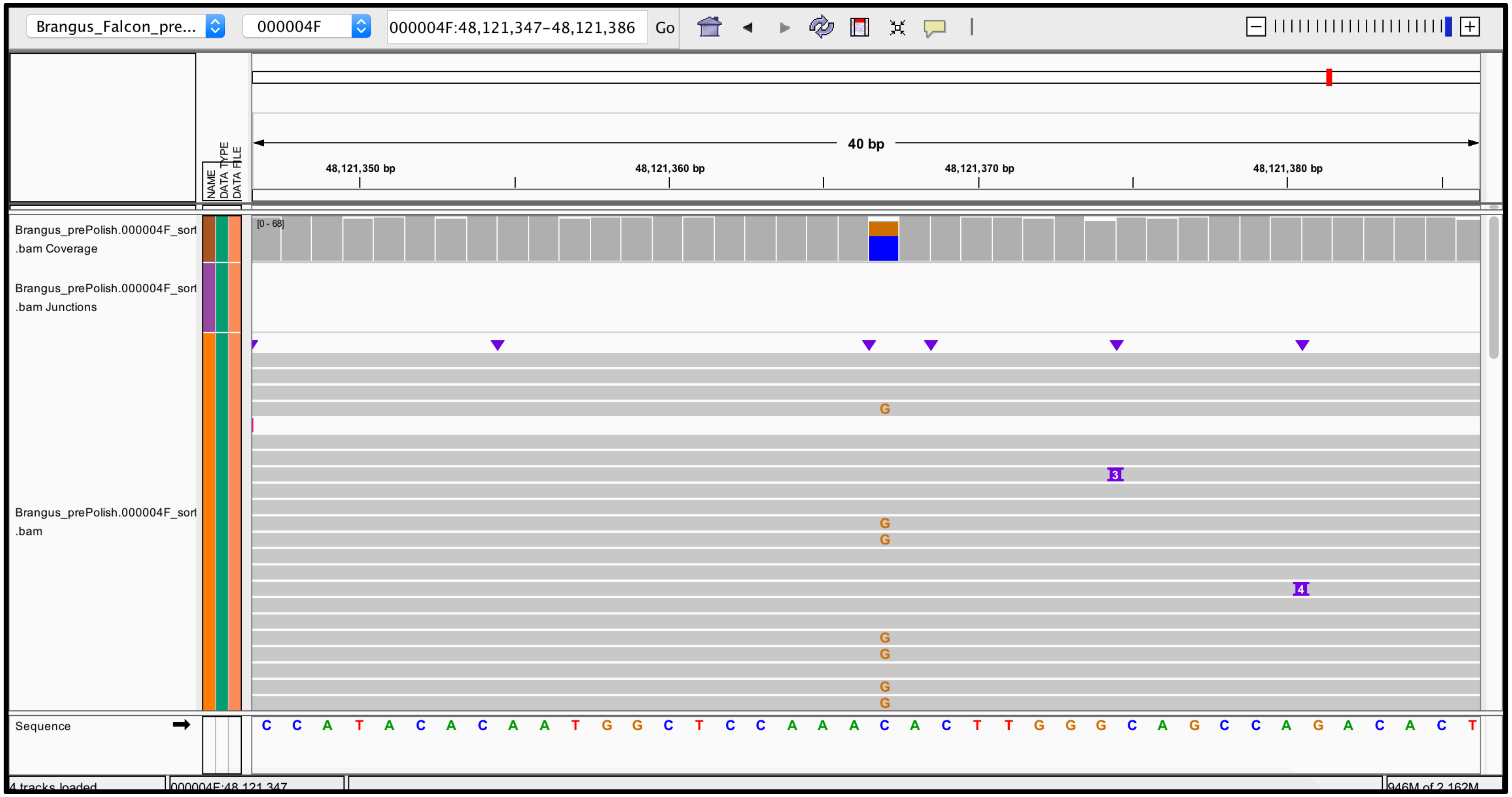


This gene (PB.1001, VPS36) contains 228 FL reads.

- 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.


Brangus_Falcon_pre... | 000004F | 000004F:48,121,347-48,121,386 | Go

NAME
DATA TYPE
DATA FILE
40 bp
48,121,350 bp
48,121,360 bp
48,121,370 bp
48,121,380 bp
Brangus_prePolish.000004F_sort .bam Coverage
[0 - 68]
Brangus_prePolish.000004F_sort .bam Junctions
Brangus_prePolish.000004F_sort .bam
G
3
G
G
4
G
G
G
G
Sequence
C C A T A C A C A A T G G C T C C A A A C A C T T G G G C A G C C A G A C A C T

4 tracks loaded | 000004F:48,121,347 | 946M of 2,162M

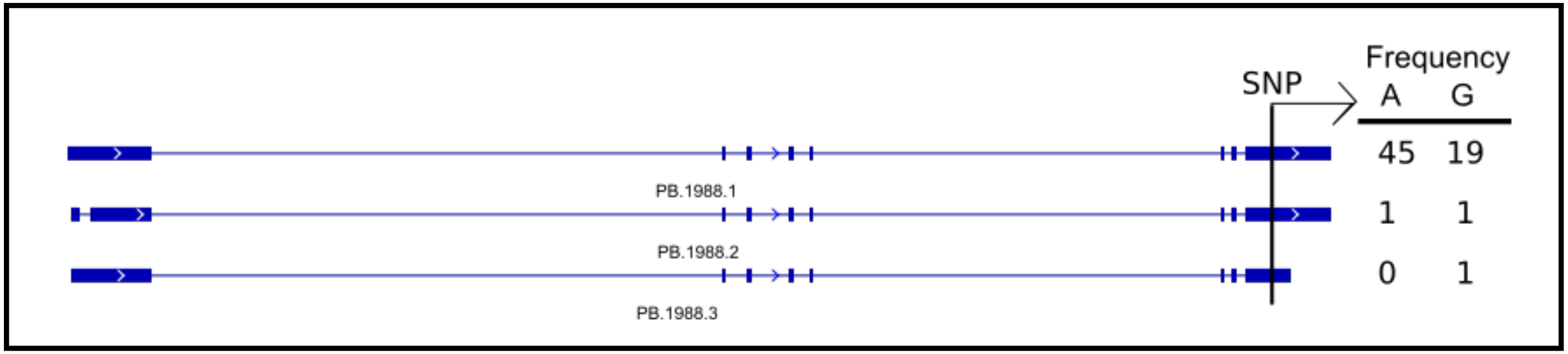
# POTENTIAL A → G RNA EDITING IN COL1A1

| CHROM | POS | REF | ALT | SNP IN GENOME? |
|---|---|---|---|---|
| **000071F** | **7663000** | **A** | **G** | **N** |
| 000071F | 7671641 | T | C | Y |

PB.8679 gene (COL1A1) contains a A → G SNP not supported by genome.

A single alternative contig (000071F_029) covers the whole region.

# POTENTIAL ALLELE IMBALANCE FOR KIF3C GENE IN BRAIN

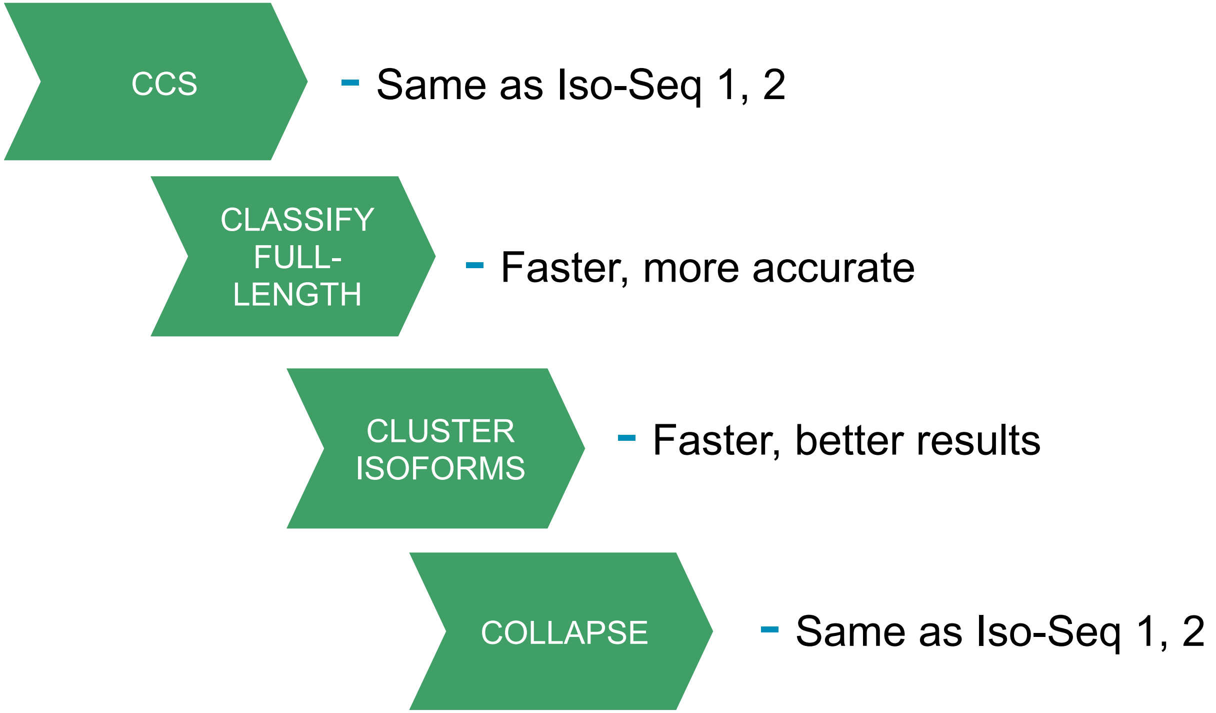


- KIF3C is observed in brain only
- The SNP is in the 3’ UTR region (A → G) and is verified by genome
- The major isoform expresses the A allele more dominantly

# Iso-Seq3 Preview

Ultra Fast + High Performance + Scalable

# ISO-SEQ3 IMPROVEMENT



CCS - Same as Iso-Seq 1, 2

CLASSIFY FULL-LENGTH - Faster, more accurate

CLUSTER ISOFORMS - Faster, better results

COLLAPSE - Same as Iso-Seq 1, 2

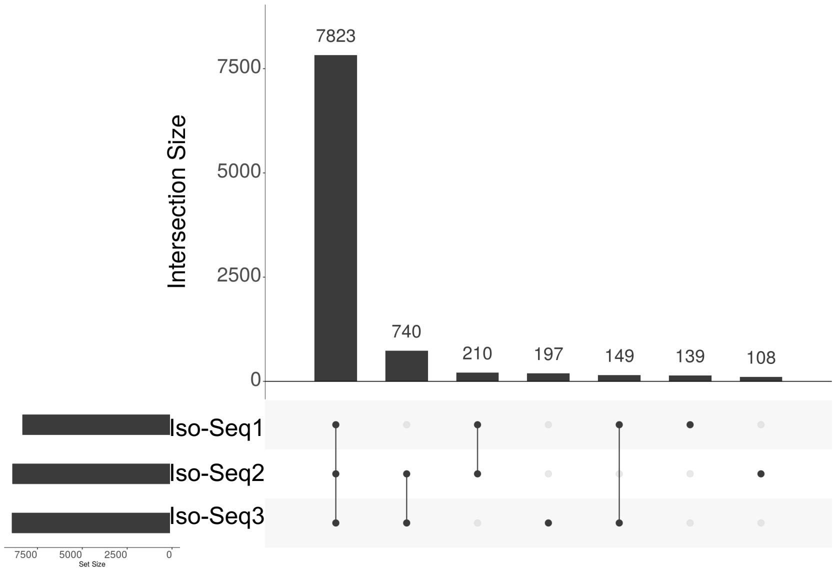
# ISO-SEQ3 IS FAST

| 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 |

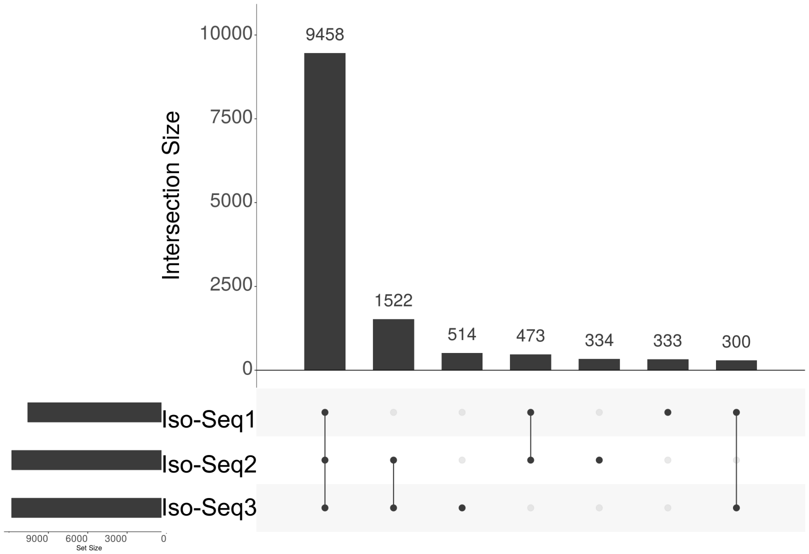
- RC0 = Universal Human Reference RNA (human) + Lexogen SIRV spike-in controls
- Not including CCS and Mapping runtime
- Computing configuration：16 CPU / node
- Tested using command line

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

## RC0 3 Cells, Known Genes Only

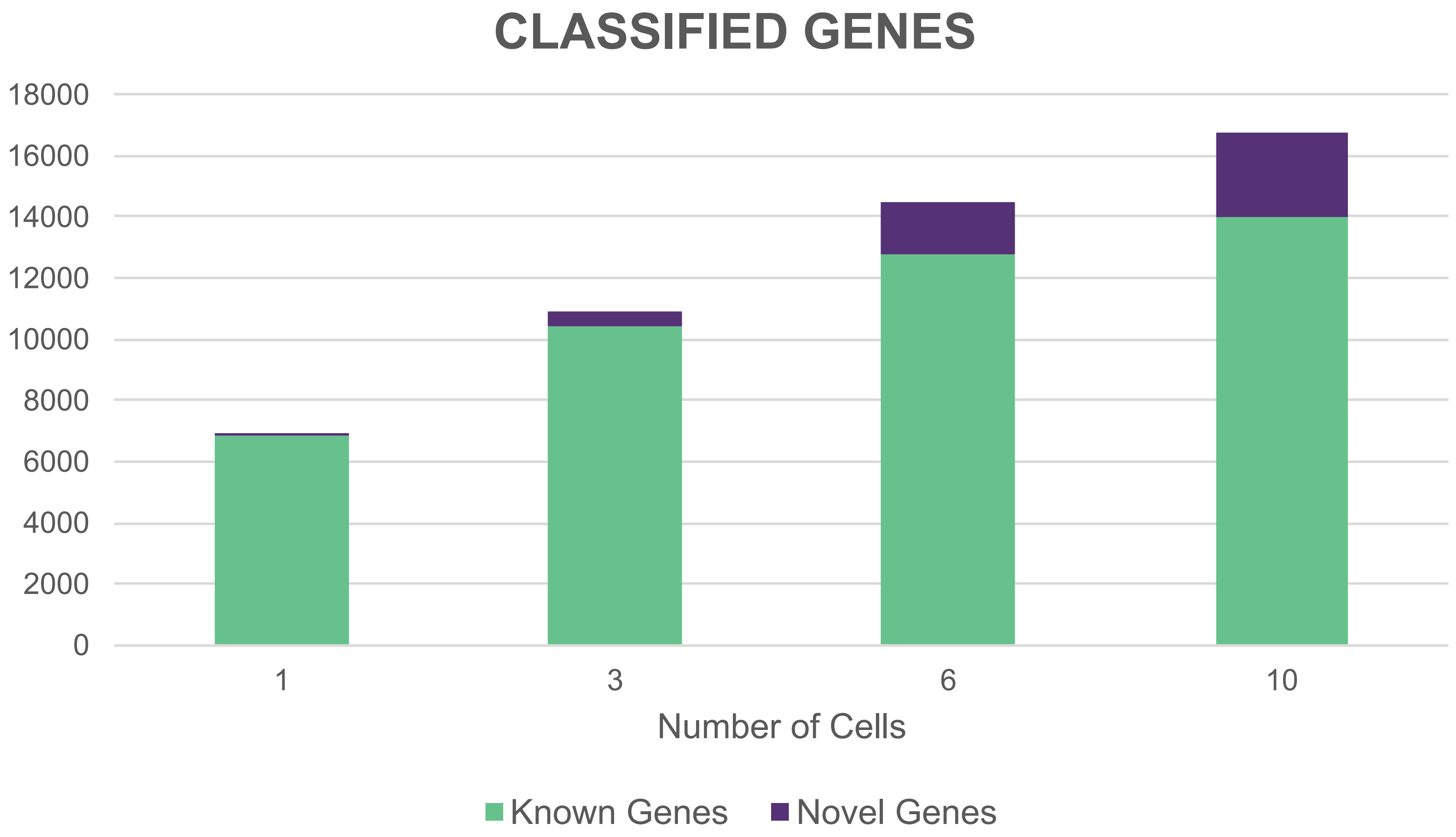


## RC0 3 Cells, Known Isoforms Only



* Only report FSM gene and isoforms

# HOW MUCH SEQUENCING IS NEEDED?



PACBIO®

www.pacb.com

For Research Use Only. Not for use in diagnostics procedures.

# Considerations for Sequencing Coverage

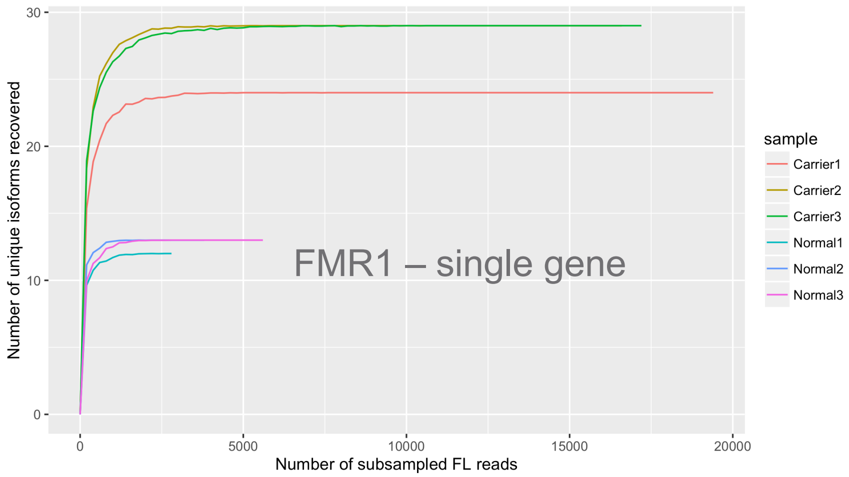
# ISO-SEQ AT SEQUEL-SCALE

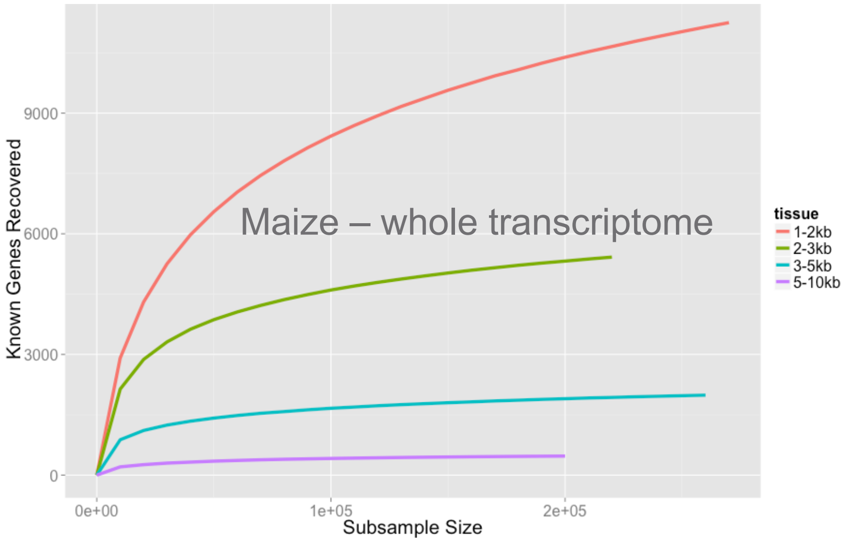
**Targeted Genes:**

- < 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 |

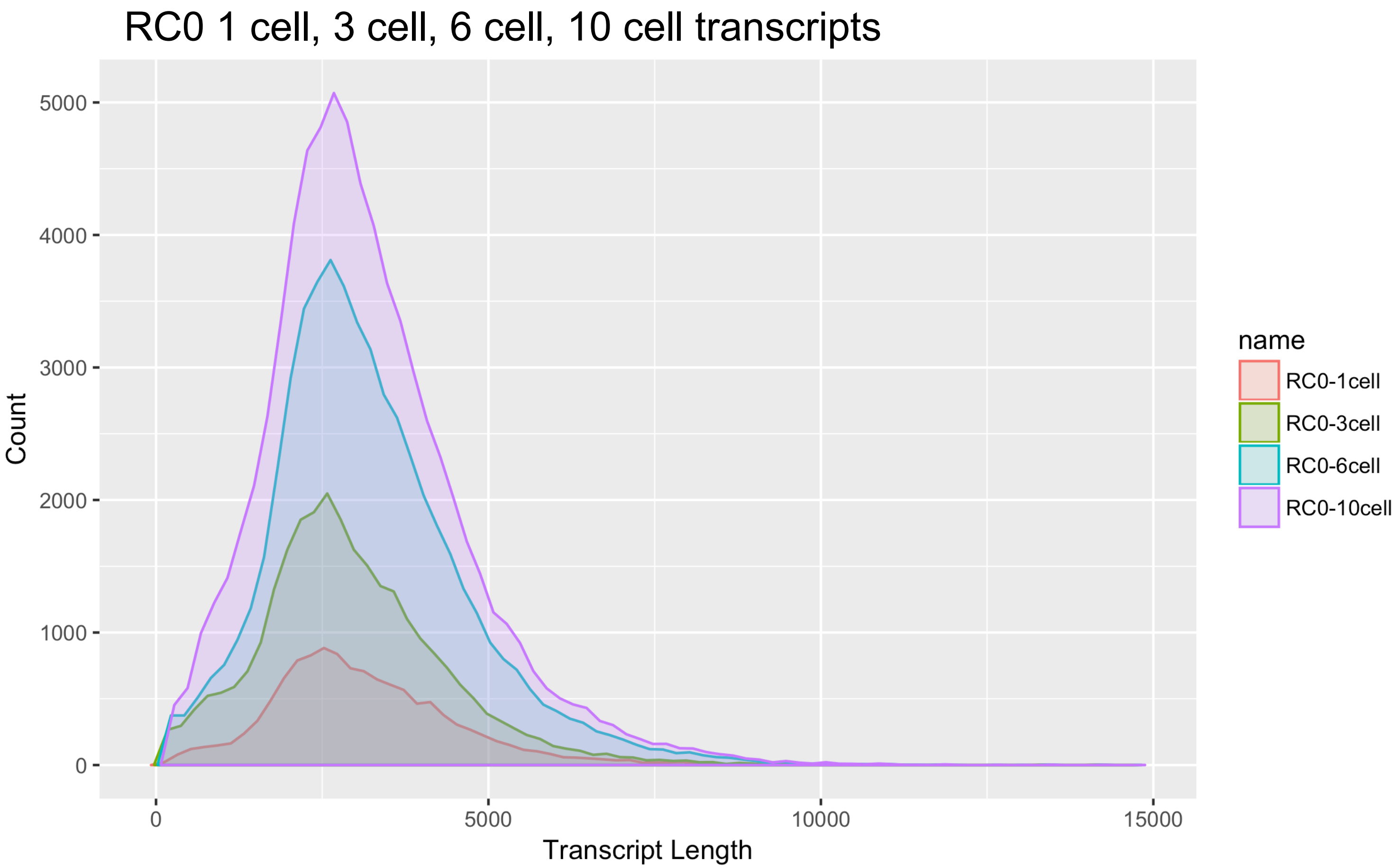
Wang et al., **Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing**, *Nat Comm* (2016)

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)

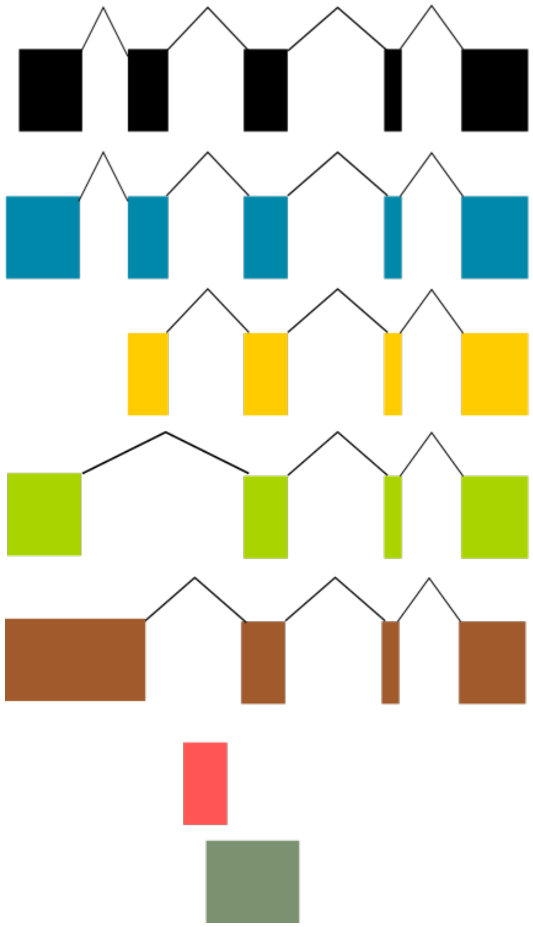
Kim et al.,**Characterization of the Rosellinia necatrix Transcriptome and Genes Related to Pathogenesis by Single-Molecule mRNA Sequencing**, *Plant Patho J* (2017)

# HUMAN TRANSCRIPTS LENGTH DISTRIBUTION



# 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

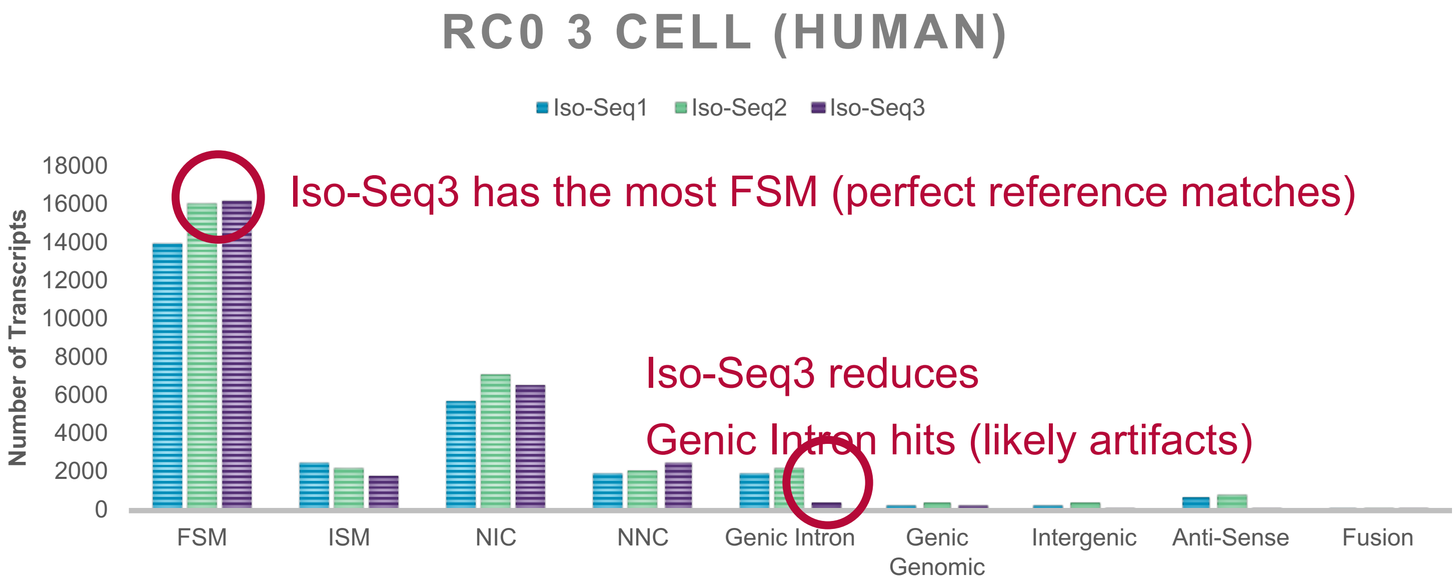
**NNC** Novel Not in Catalog, novel isoforms using novel junctions.

**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: HUMAN

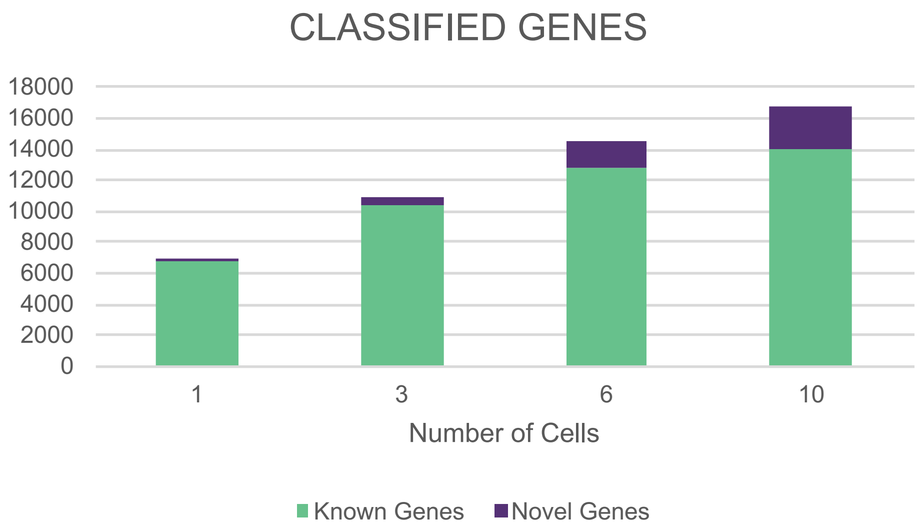


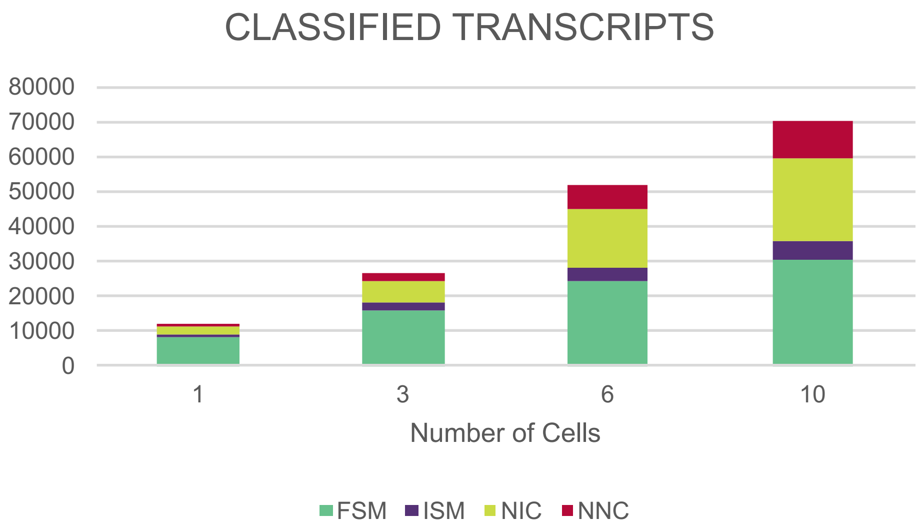
Iso-Seq3 has the most FSM (perfect reference matches)

Iso-Seq3 reduces Genic Intron hits (likely artifacts)

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

# HOW MUCH SEQUENCING IS NEEDED?


CLASSIFIED GENES
Number of Cells
Known Genes
Novel Genes


CLASSIFIED TRANSCRIPTS
Number of Cells
FSM
ISM
NIC
NNC