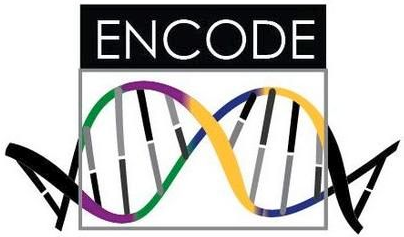
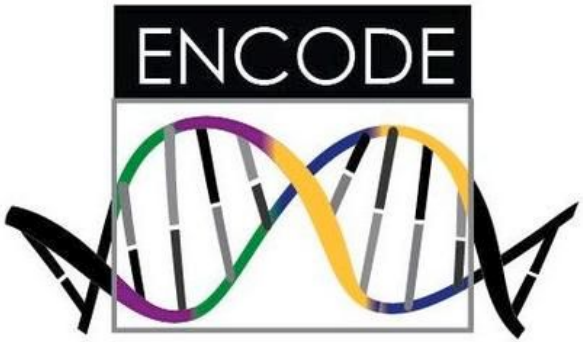
# Developing technology-agnostic tools for analyzing long read transcriptome data


Dana Wyman
Mortazavi Lab
Lab Meeting, 4/19/2018



# Goals of our lab within the ENCODE4 project

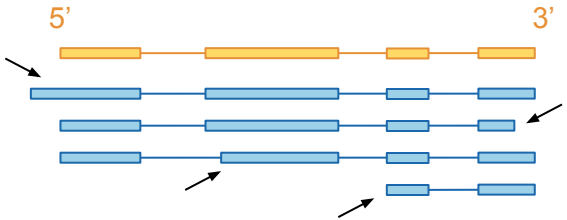
1. Use long read sequencing to ask which gene isoforms are expressed in different cell types
2. For each cell type, provide a high-quality annotation of both known and novel isoforms that are expressed
    a. Serve as a reference for improved short-read quantitation of gene expression



As it turns out, comparing transcripts can be complicated

# As it turns out, comparing transcripts can be complicated

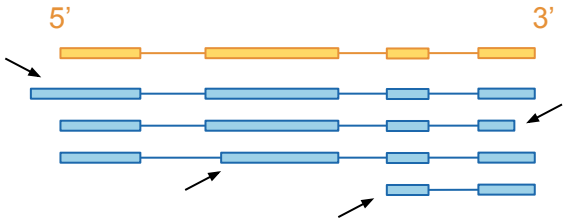
How do we decide whether two transcripts are the same?




Exact match approach tends to be too strict

# As it turns out, comparing transcripts can be complicated

How do we decide whether two transcripts are the same?



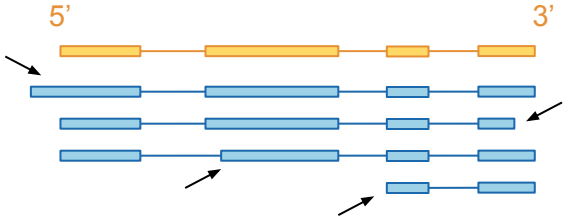

Exact match approach tends to be too strict

How do we track novel transcripts across different datasets?

GeneA-NovelTranscript1  GeneA-NovelTranscript2

# As it turns out, comparing transcripts can be complicated

How do we decide whether two transcripts are the same?




Exact match approach tends to be too strict

How do we distinguish novel biology from artifacts?


Degraded RNA or new isoform?

How do we track novel transcripts across different datasets?

GeneA-NovelTranscript1 ⟷ GeneA-NovelTranscript2

# How do we compare transcriptomes from different long read platforms?

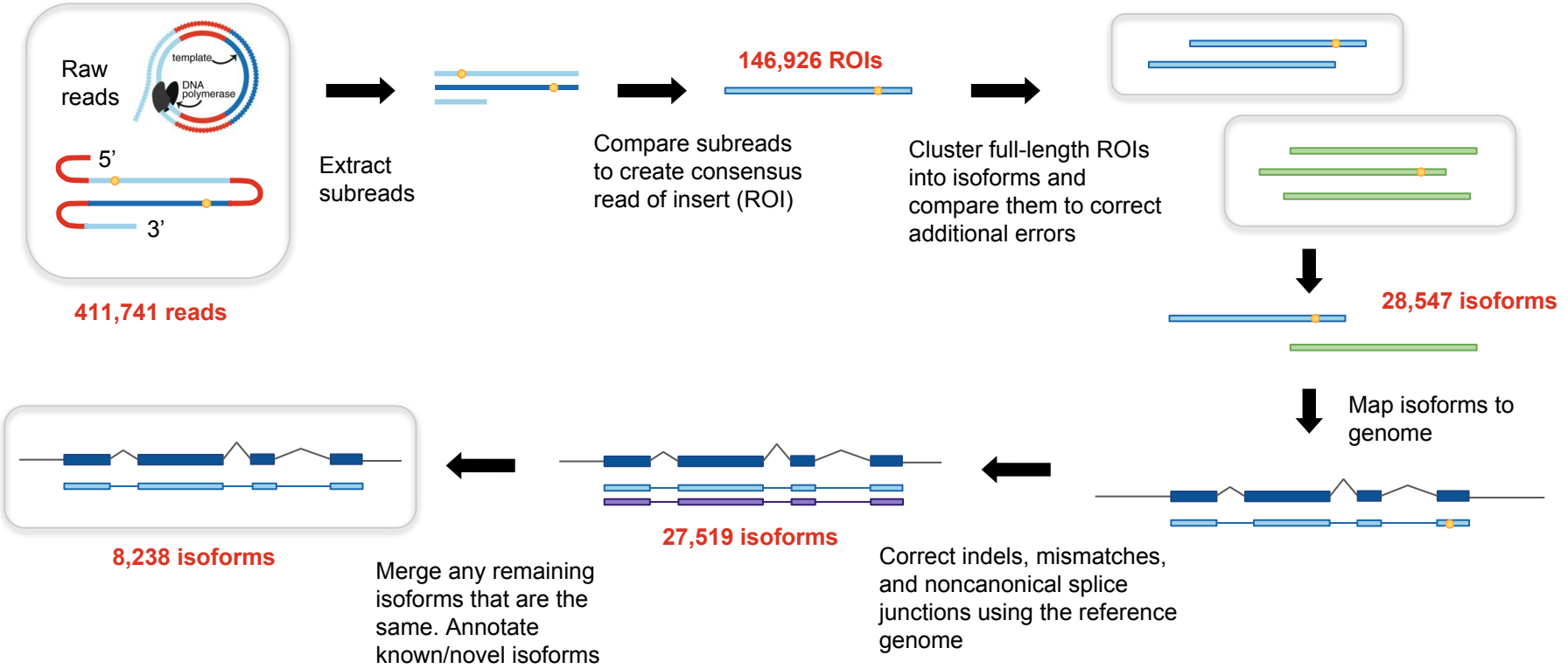
PacBio RSII

PacBio Sequel

Oxford Nanopore MinION

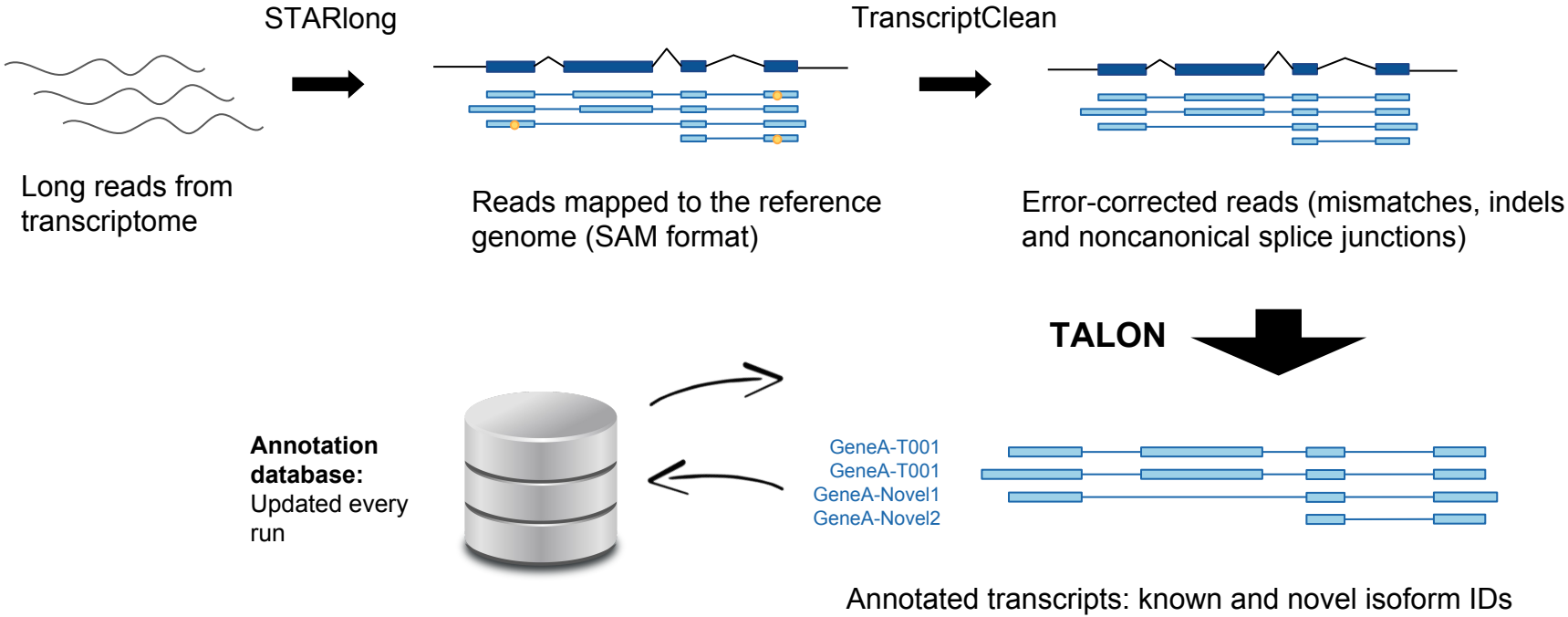
Pacific Biosciences, Oxford Nanopore Technologies

# Current PacBio pipeline is long and throws out a lot of data



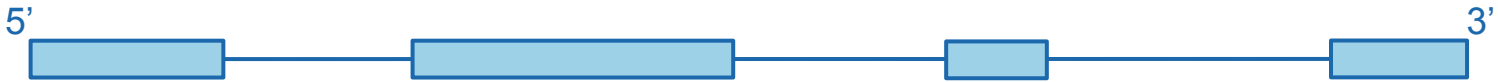

Counts from PB36 Dataset

# TALON: Technology-Agnostic Long Read Analysis Pipeline



# Alpha version of TALON

## Looking for isoform matches for a query transcript




Let’s say we want to find the best annotated isoform match for this mapped transcript.

# Alpha version of TALON

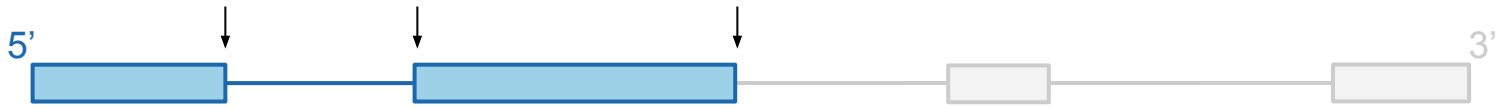
## Looking for isoform matches for a query transcript


Start with the first exon.

Find all annotation transcripts that contain this junction.
**Variation allowed at 5’ end.**

# Alpha version of TALON

## Looking for isoform matches for a query transcript



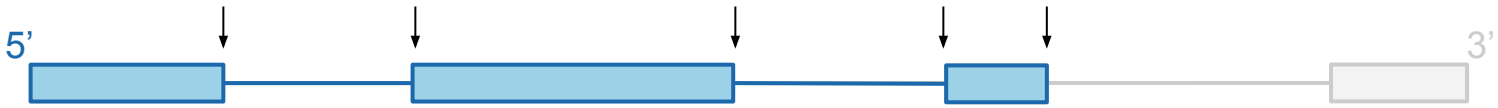

Find all annotation transcripts that contain this junction. **Variation allowed at 5’ end.**

∩

Find all annotation transcripts that contain these exact junctions.

# Alpha version of TALON

## Looking for isoform matches for a query transcript




Find all annotation transcripts that contain this junction. **Variation allowed at 5’ end.**

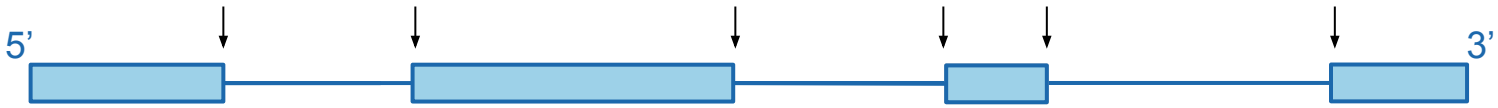
∩

Find all annotation transcripts that contain these exact junctions.

∩ . . . . . .

# Alpha version of TALON

## Looking for isoform matches for a query transcript




Find all annotation transcripts that contain this junction. **Variation allowed at 5’ end.**

∩

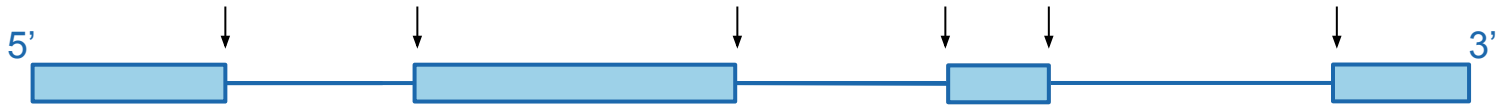
Find all annotation transcripts that contain these exact junctions.

∩ . . . . . ∩

Find all annotation transcripts that contain this junction. **Variation allowed at 3’ end.**

# Alpha version of TALON

## Looking for isoform matches for a query transcript




Find all annotation transcripts that contain this junction. **Variation allowed at 5’ end.**

∩

Find all annotation transcripts that contain these exact junctions.

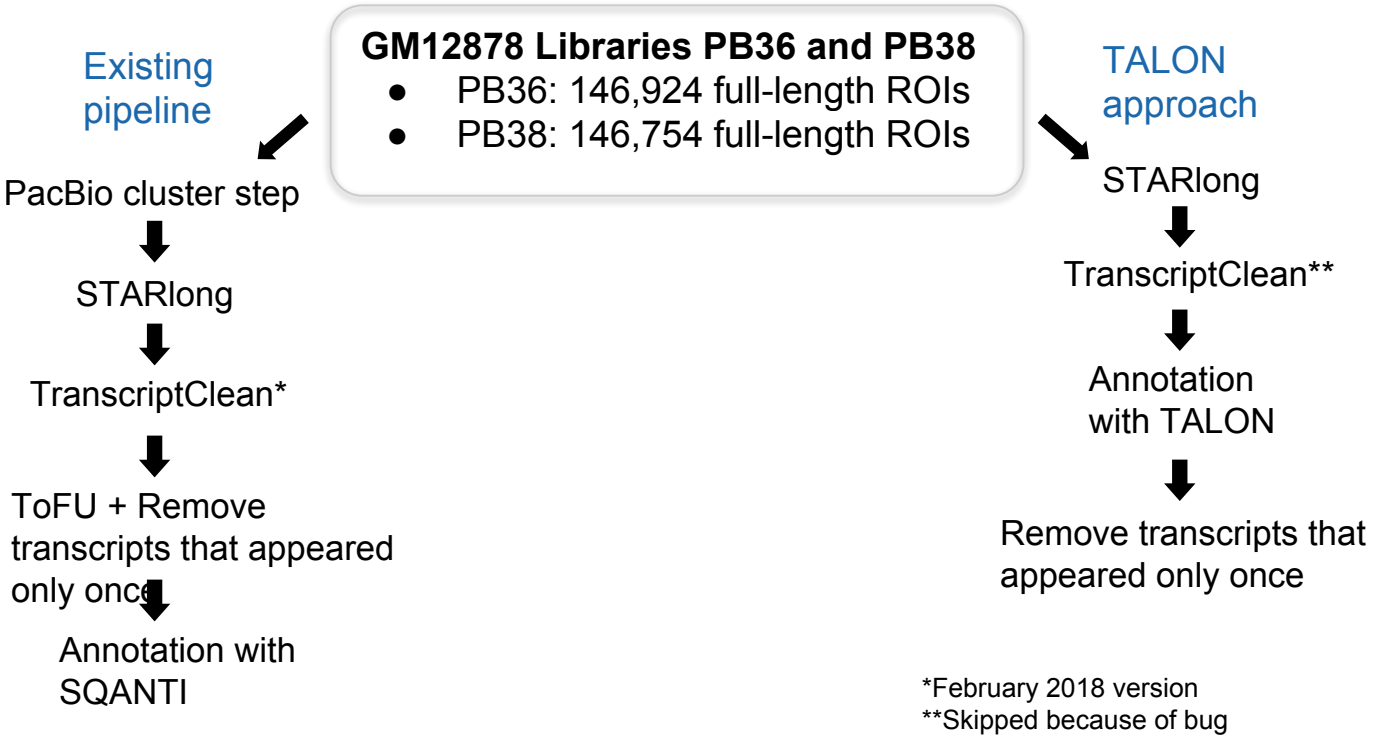
∩ . . . . . ∩

Find all annotation transcripts that contain this junction. **Variation allowed at 3’ end.**

Possibilities at the end:

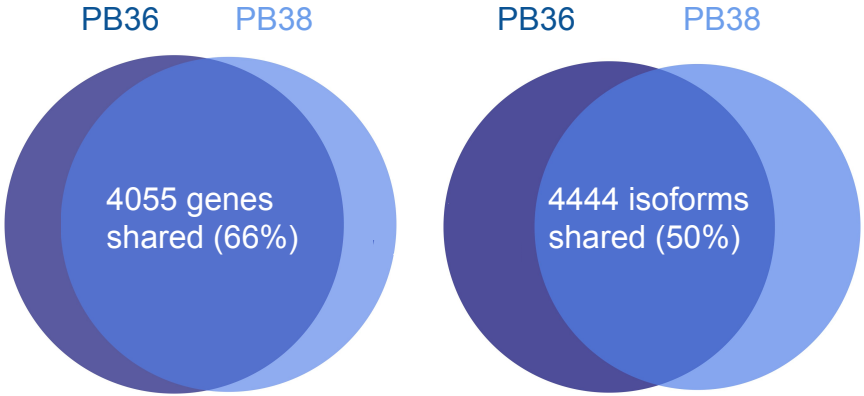
- Final set contains ≥ 1 isoform → Known transcript
- Final set is empty but others are not → Novel transcript of known gene
- All sets are empty → Novel transcript that may belong to a known or novel gene
  - Not fully implemented yet

# Comparing TALON performance to the existing PacBio pipeline on biological replicate data



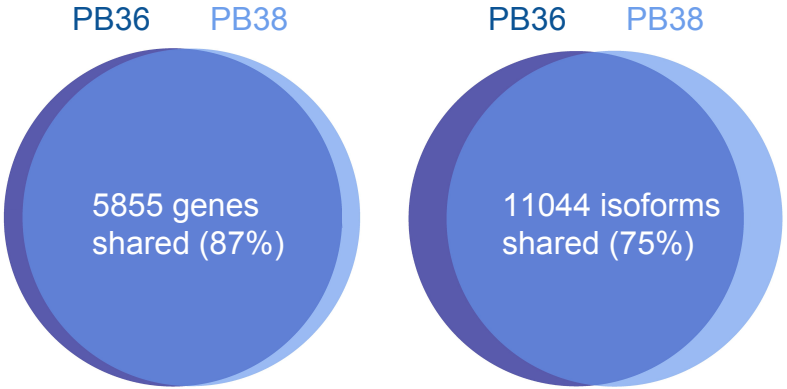
# TALON detects more genes and transcripts than the existing pipeline and is more reproducible

## Existing Pipeline



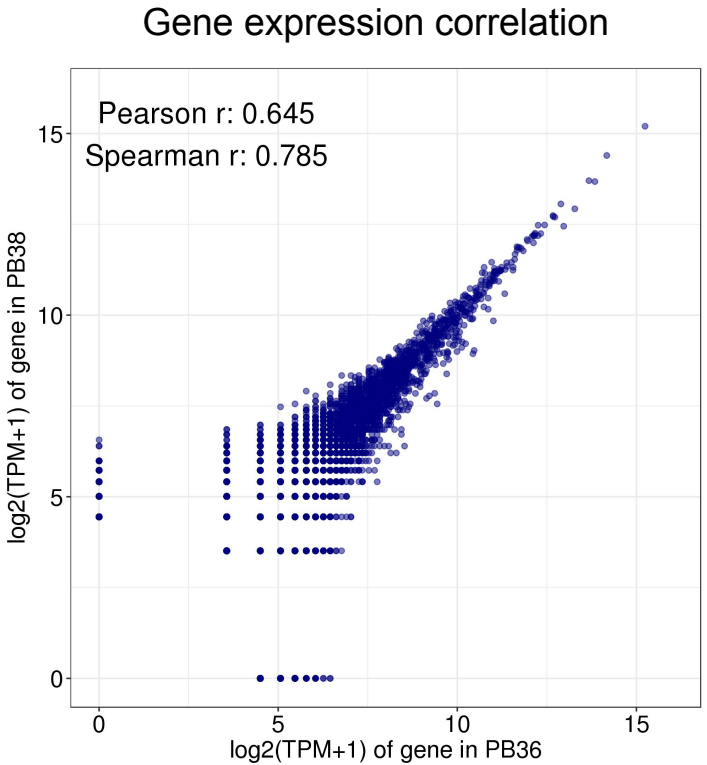

- Detected 6,134 genes
- Detected 8,983 isoforms

## TALON

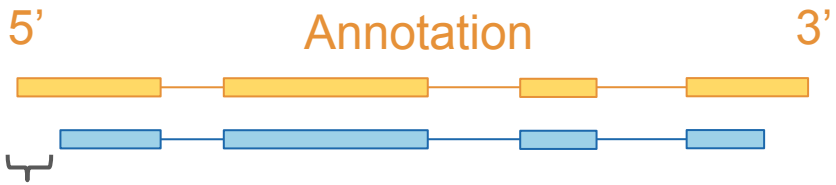



- Detected 6,723 genes
- Detected 14,789 isoforms

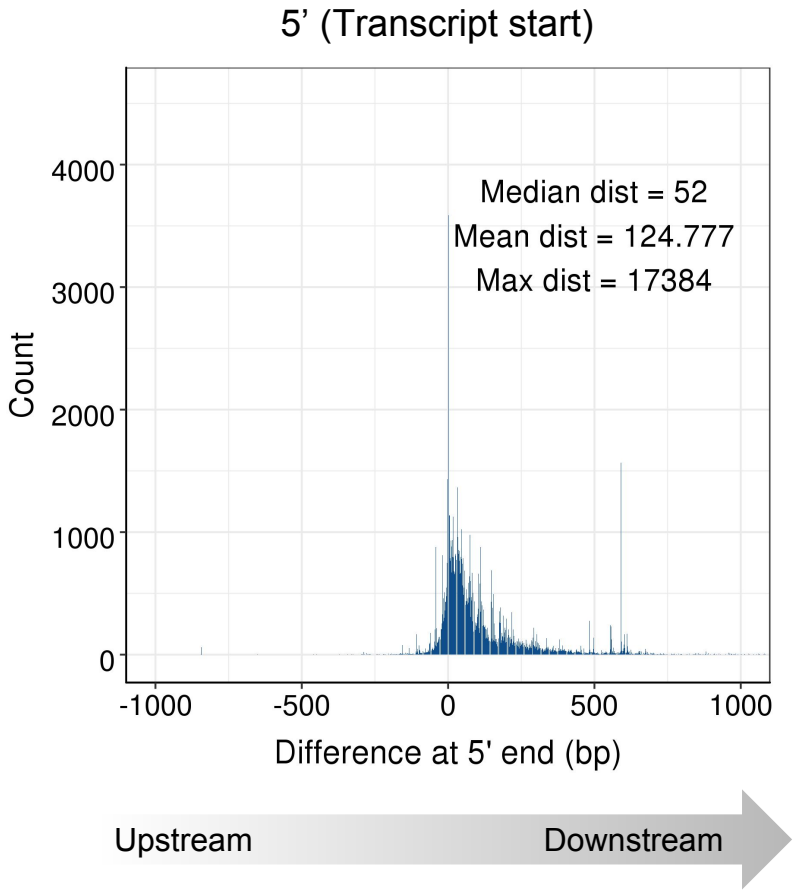
# TALON-annotated GM12878 replicates show strongly correlated gene expression



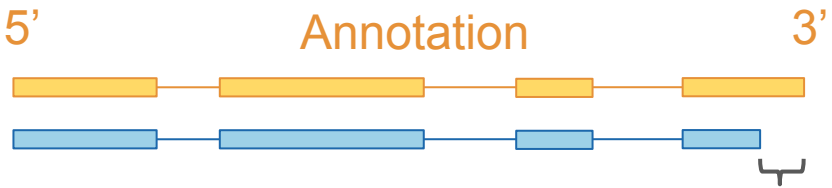
# Known transcripts display 5’ and 3’ end variation



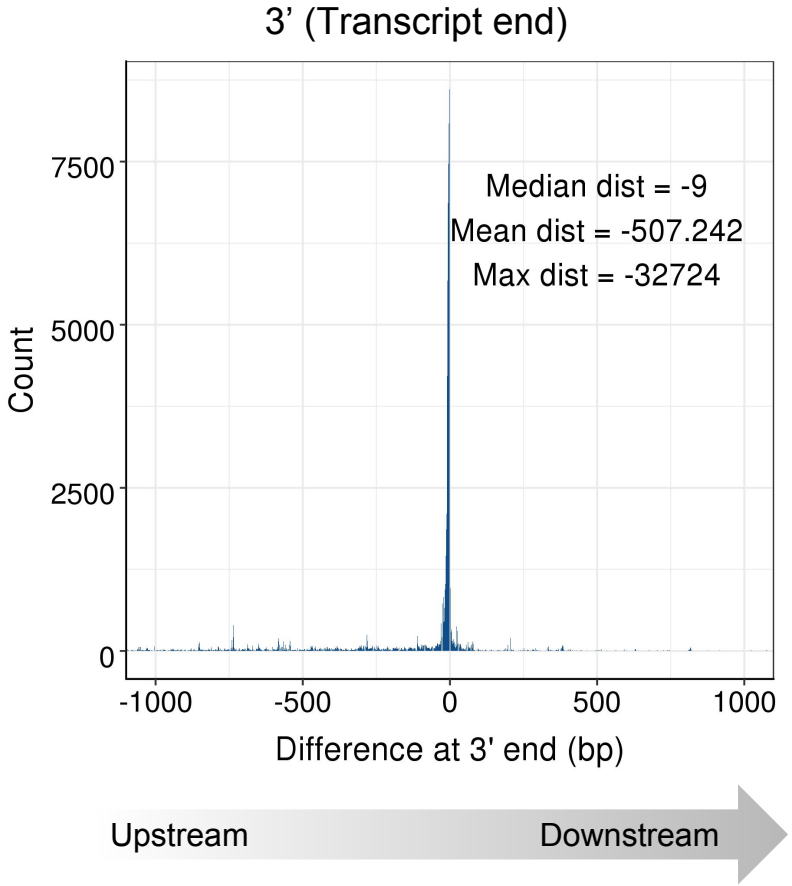

Exons in sequenced transcript match a known isoform, but there is a difference at the 5’ end.



# Known transcripts display 5’ and 3’ end variation




Exons in sequenced transcript match a known isoform, but there is a difference at the 3’ end.

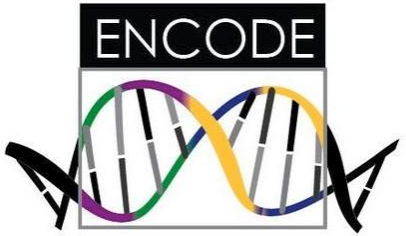


# (Near) Future Directions

- Implement a structure to track our transcript annotations over time so that new TALON runs are automatically compared to previous runs
- Refine exon matching approach to provide additional information about partial matches to known transcripts
- Perform a more comprehensive comparison between TALON results and the old PacBio pipeline
- Run TALON on lots of data!
    - Compare additional cell line datasets
    - Compare PacBio to Oxford Nanopore

# Acknowledgements


- **Mortazavi Lab**
  - Gabriela Balderrama Gutierrez
  - Dr. Weihua “Benny” Zeng
  - Shan “Mandy” Jiang
  - Camden Jansen
  - Kate Williams
  - Rabi Murad
  - Lorrayne Serra
  - Christina Wilcox
  - Klebea Carvalho
  - Sorena Rahmanian
  - Xinyi “Savanna” Ma

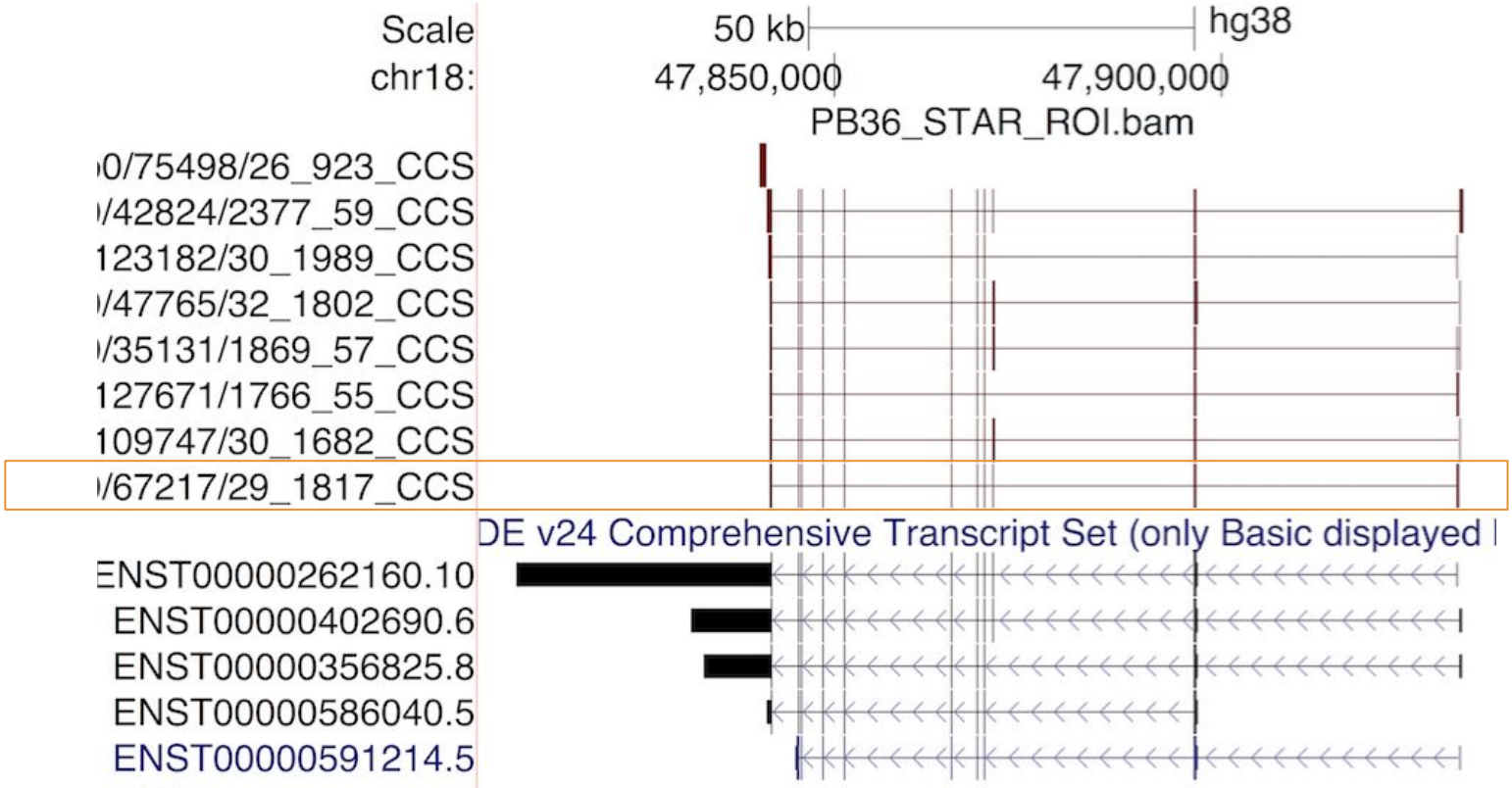





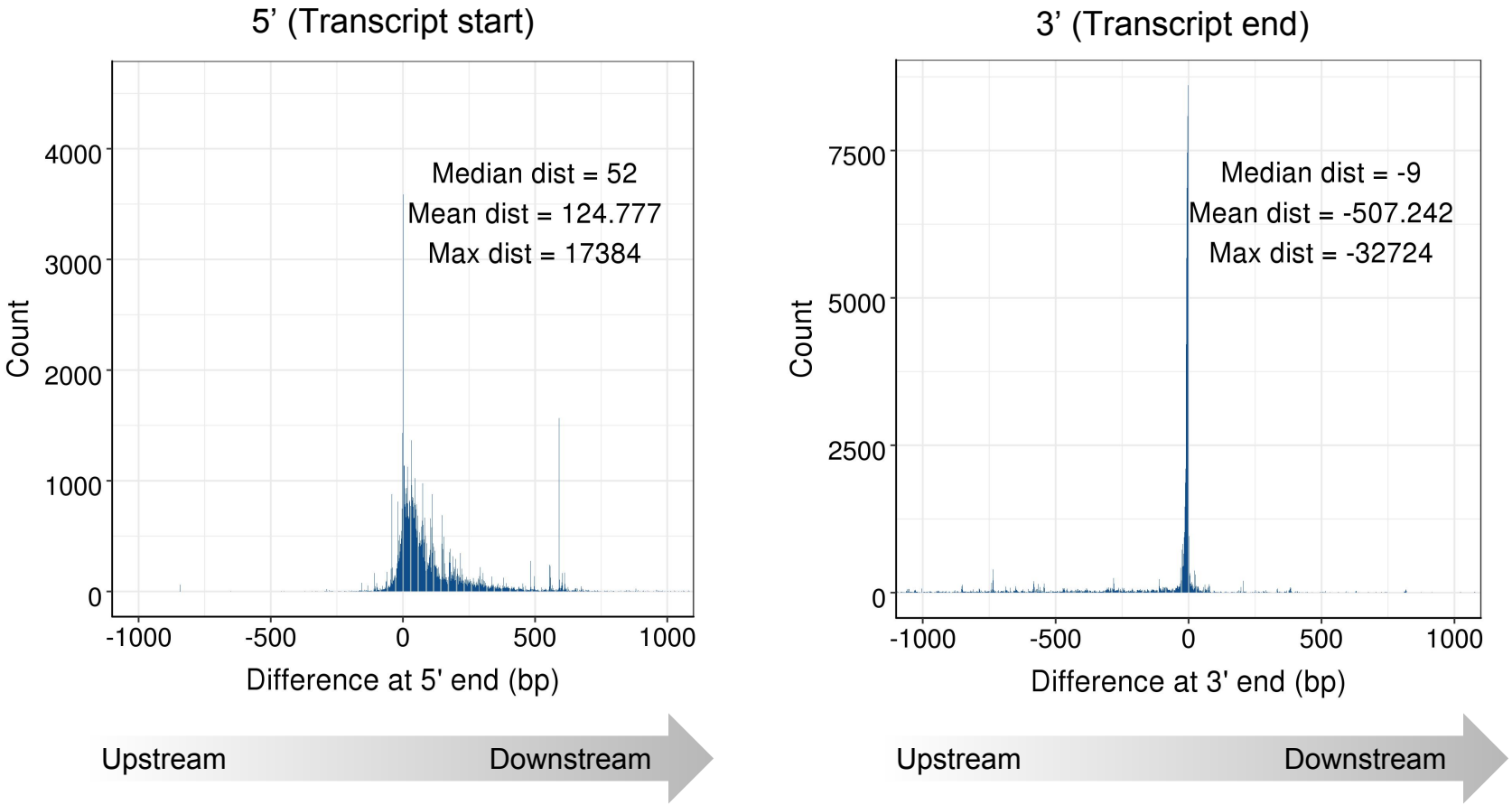

Center for Complex Biological Systems
University of California, Irvine

# Extra slides

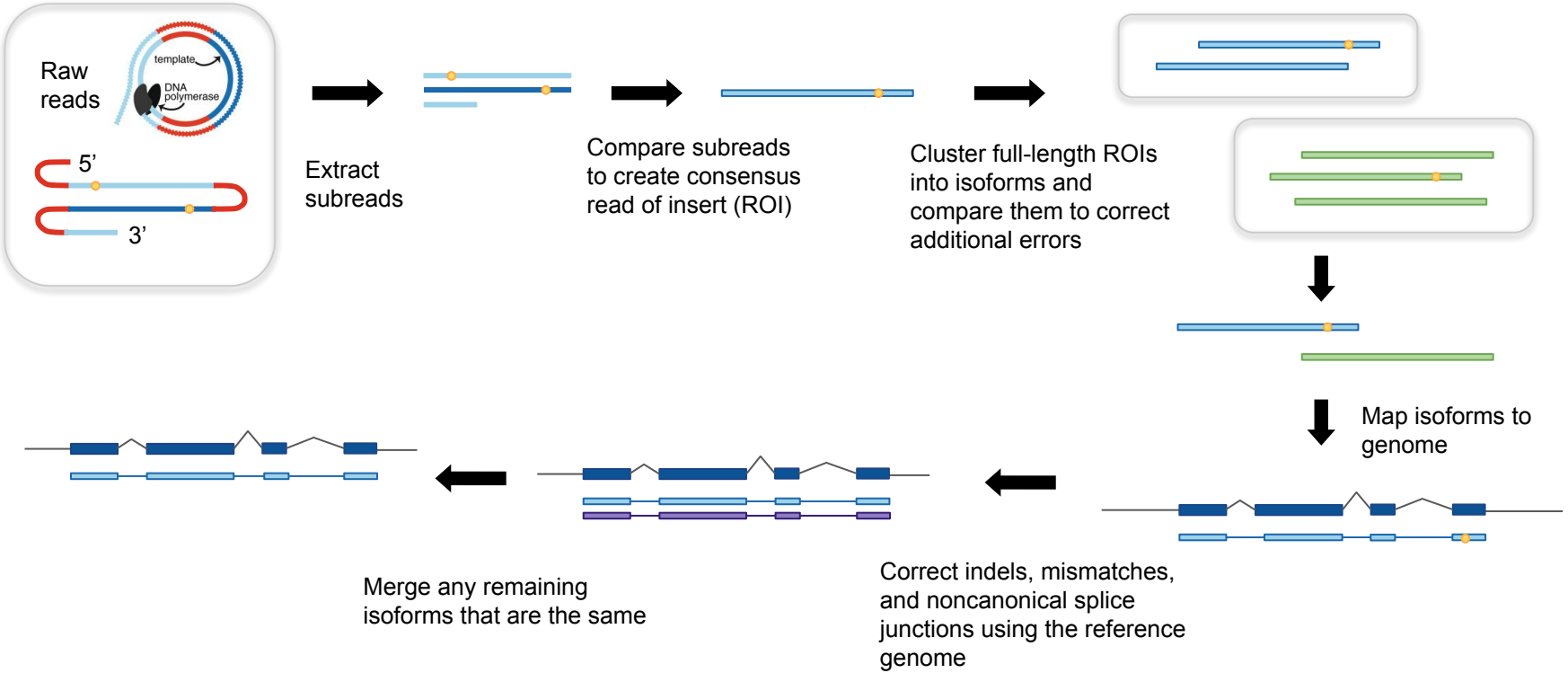
# Example of a transcript with a large 3’ difference from the annotation



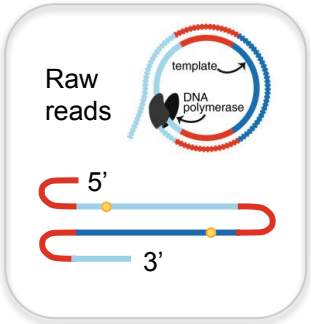
# Known transcripts show extensive 5’ and 3’ end variation



# Processing raw PacBio RSII reads into isoforms



# New Pipeline



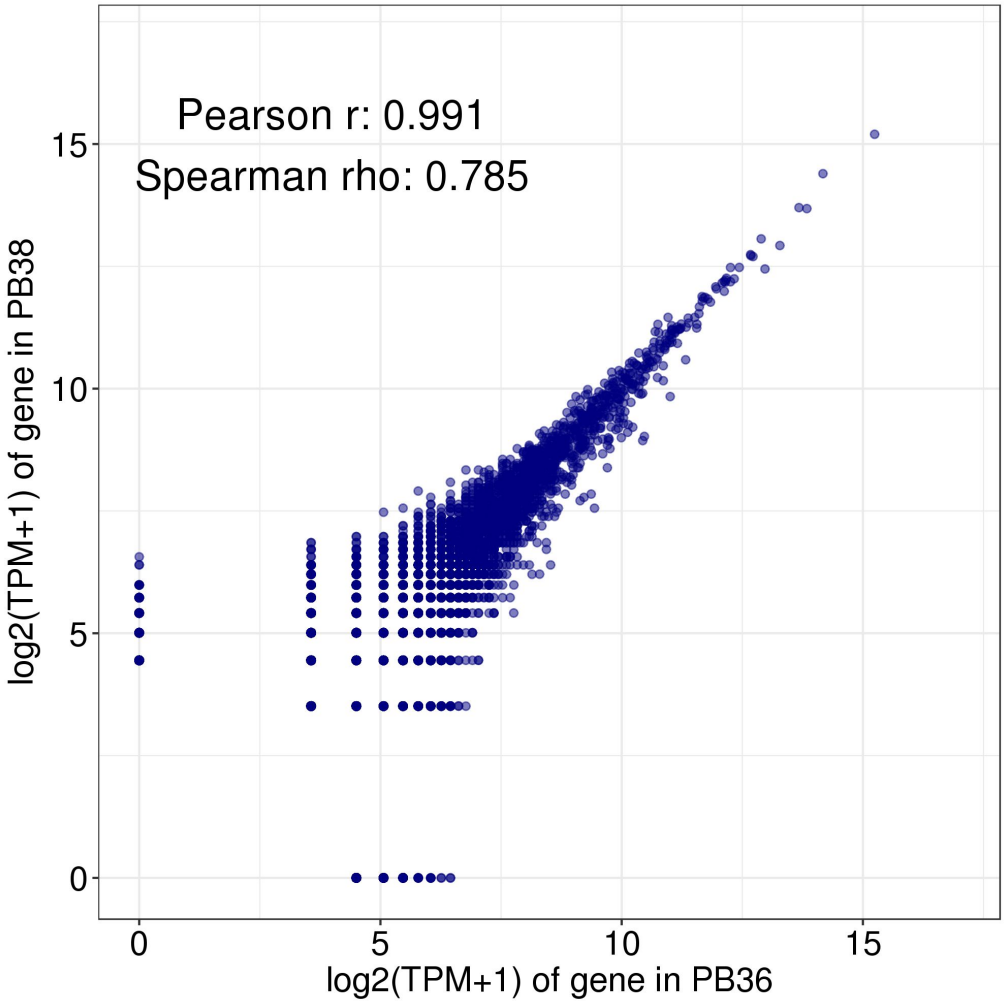

Extract subreads

Compare subreads to create consensus read of insert (ROI)

Map isoforms to reference genome

**TranscriptClean:** Correct indels, mismatches, and noncanonical splice junctions using the reference genome

**TALON:** Annotate transcripts to assign gene and isoform identity

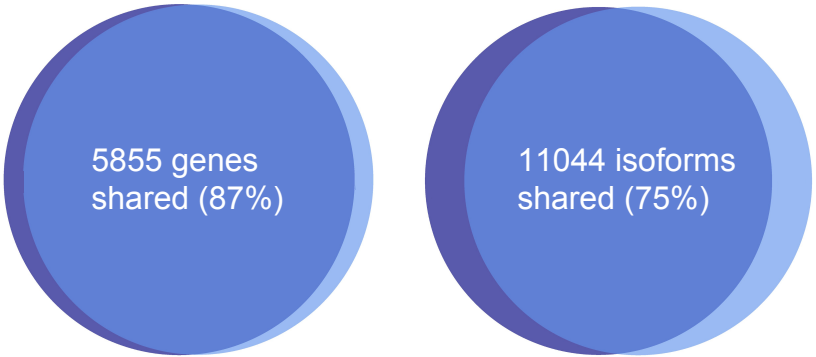


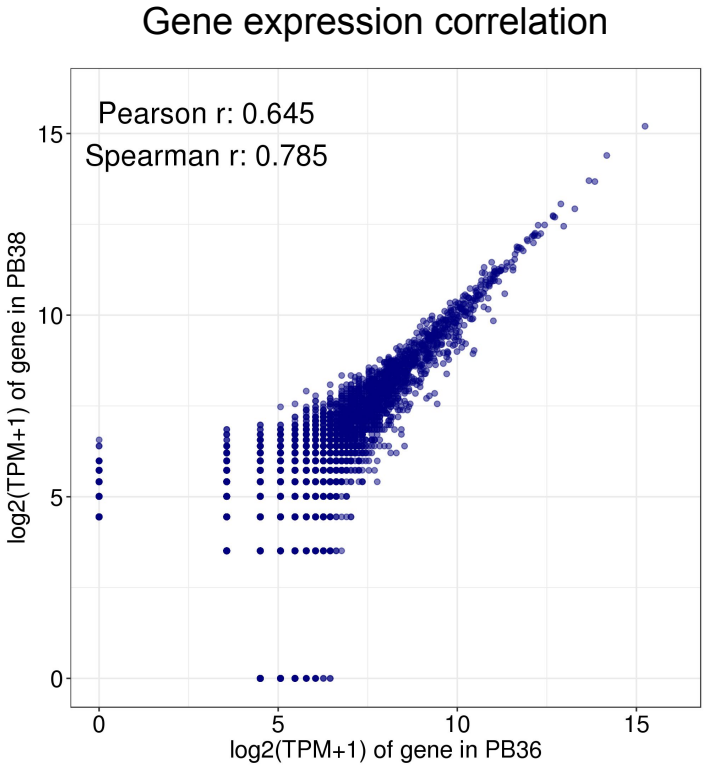

Gene expression correlation when run directly on the TPMs rather than log2(TPM+1)

# TALON Results: Transcripts detected in GM12878 biological replicates are largely reproducible

After filtering and combining the replicates:

- Detected 6,723 genes
- Detected 14,789 transcripts
  - 7123 known
  - 7666 novel



# PB36 TranscriptClean results