

# RNA-seq解析パイプライン： *de novo*

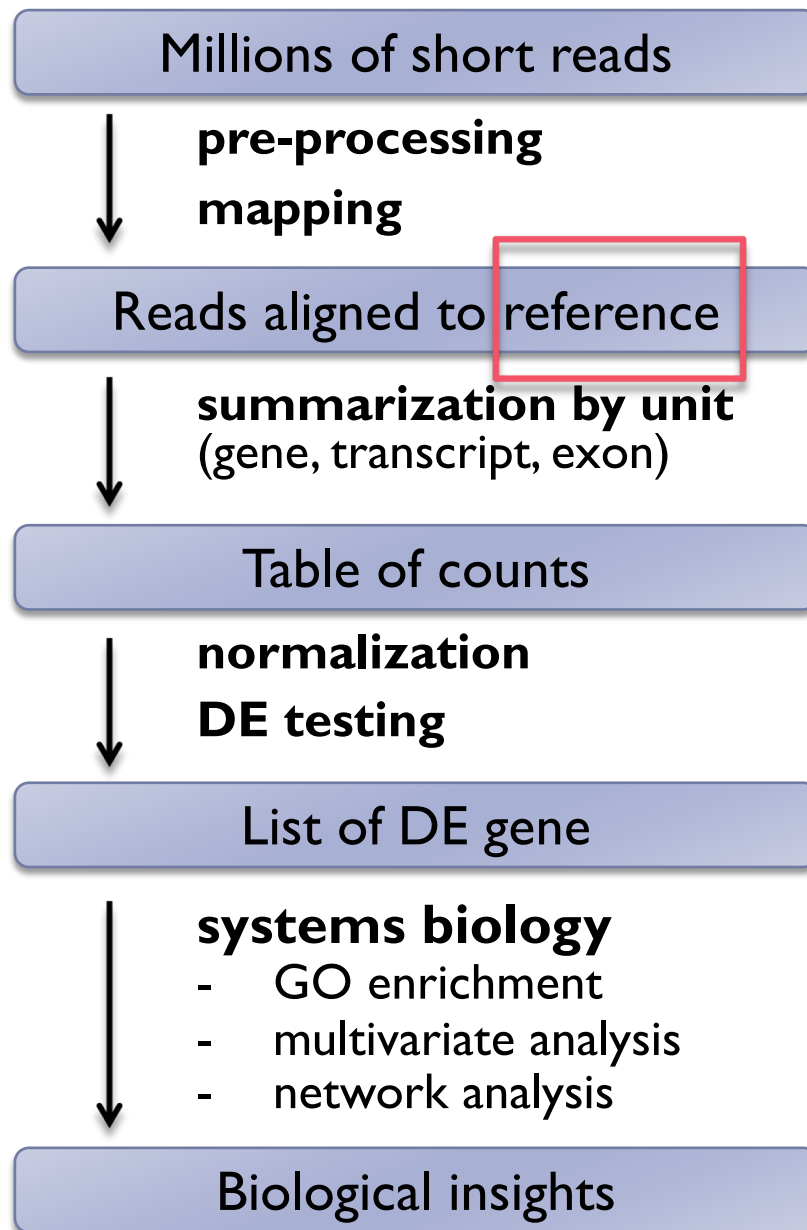
Shuji Shigenobu

重信 秀治

基礎生物学的研究所  
生物機能解析センター

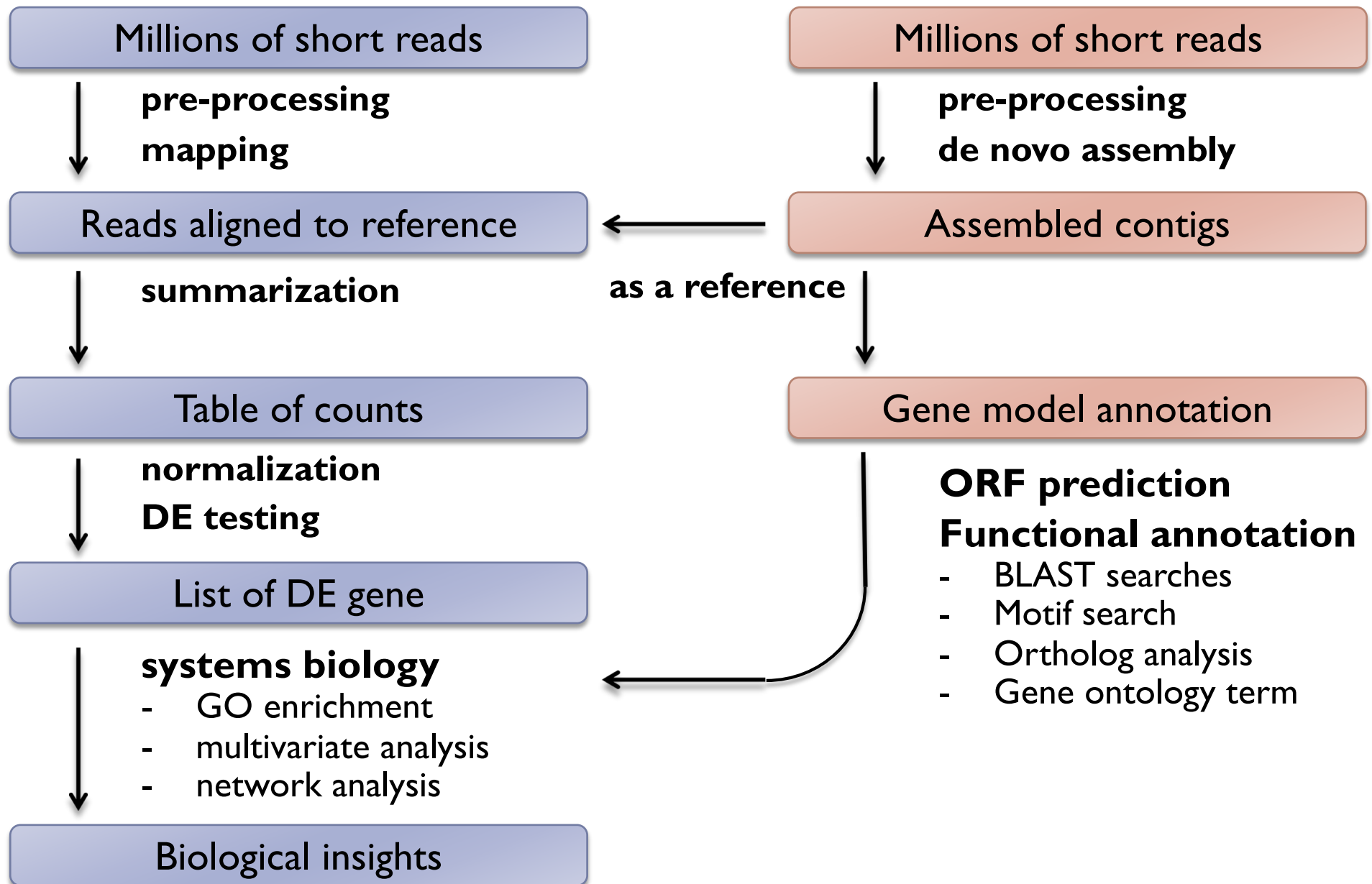


# *de novo* RNA-seq



1. **Build** reference
2. **Characterize** reference

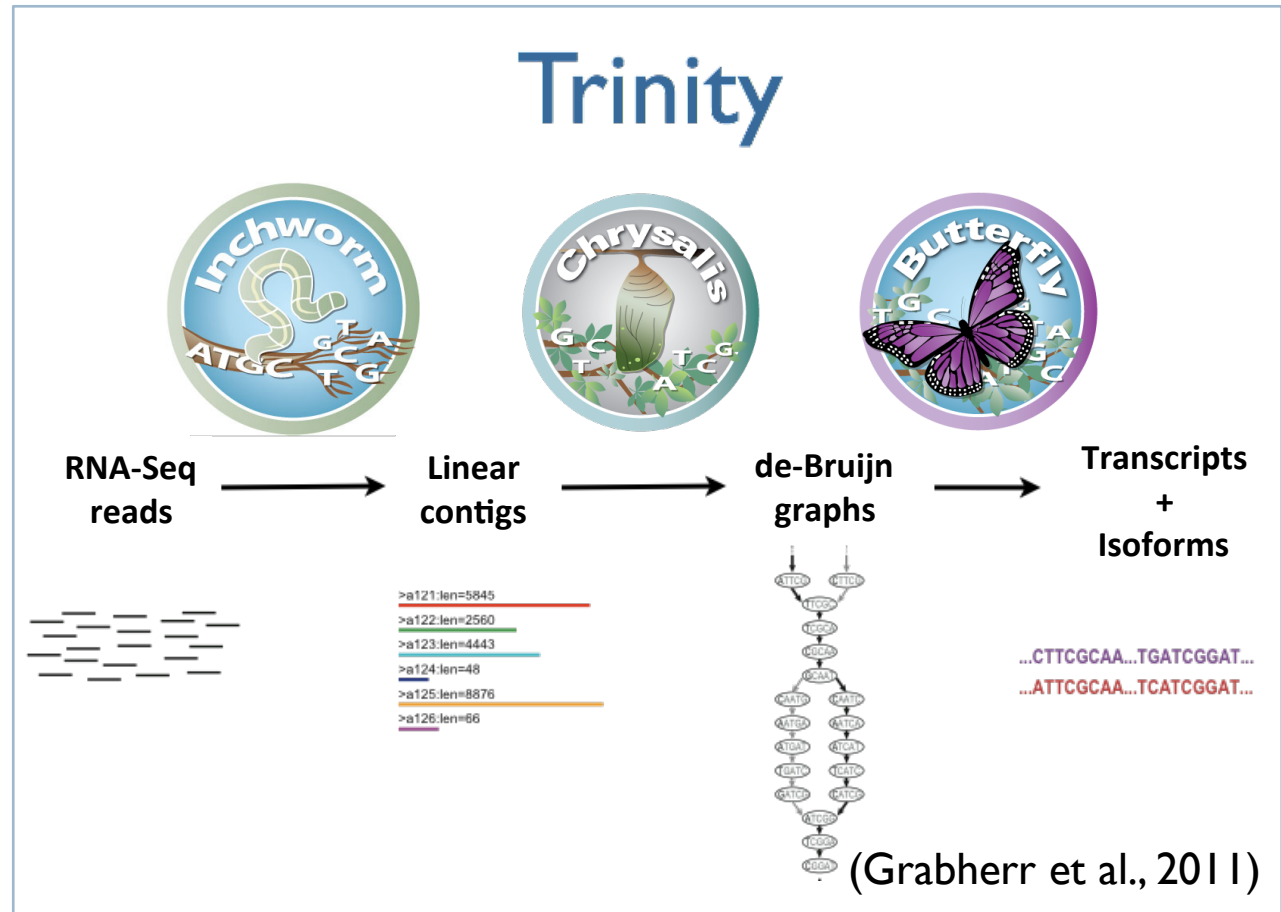
# RNA-seq analysis pipeline (*de novo* strategy)



# *de novo* assemblers of RNA-seq

*De novo* assemblers use reads to assemble transcripts directly, which does not depend on a reference genome.

- ▶ Trinity
- ▶ Oases
- ▶ TransAbyss
- ▶ ...



# Home

Brian Haas edited this page on Nov 1, 2017 · 35 revisions

<https://github.com/trinityrnaseq/trinityrnaseq/wiki>

## RNA-Seq De novo Assembly Using Trinity

► Pages 30



### Quick Guide for the Impatient

Trinity assembles transcript sequences from Illumina RNA-Seq data.

Download Trinity [here](#).

- [Trinity Wiki Home](#)
- [Installing Trinity](#)
  - [Trinity Computing Requirements](#)
  - [Accessing Trinity on Publicly Available Compute Resources](#)
  - [Run Trinity using Docker](#)
- [Running Trinity](#)
  - [Genome Guided Trinity Transcriptome Assembly](#)
  - [Gene Structure Annotation of Genomes](#)
- [Trinity process and resource monitoring](#)
  - [Monitoring Progress During a Trinity Run](#)
  - [Examining Resource Usage at the End of a Trinity Run](#)

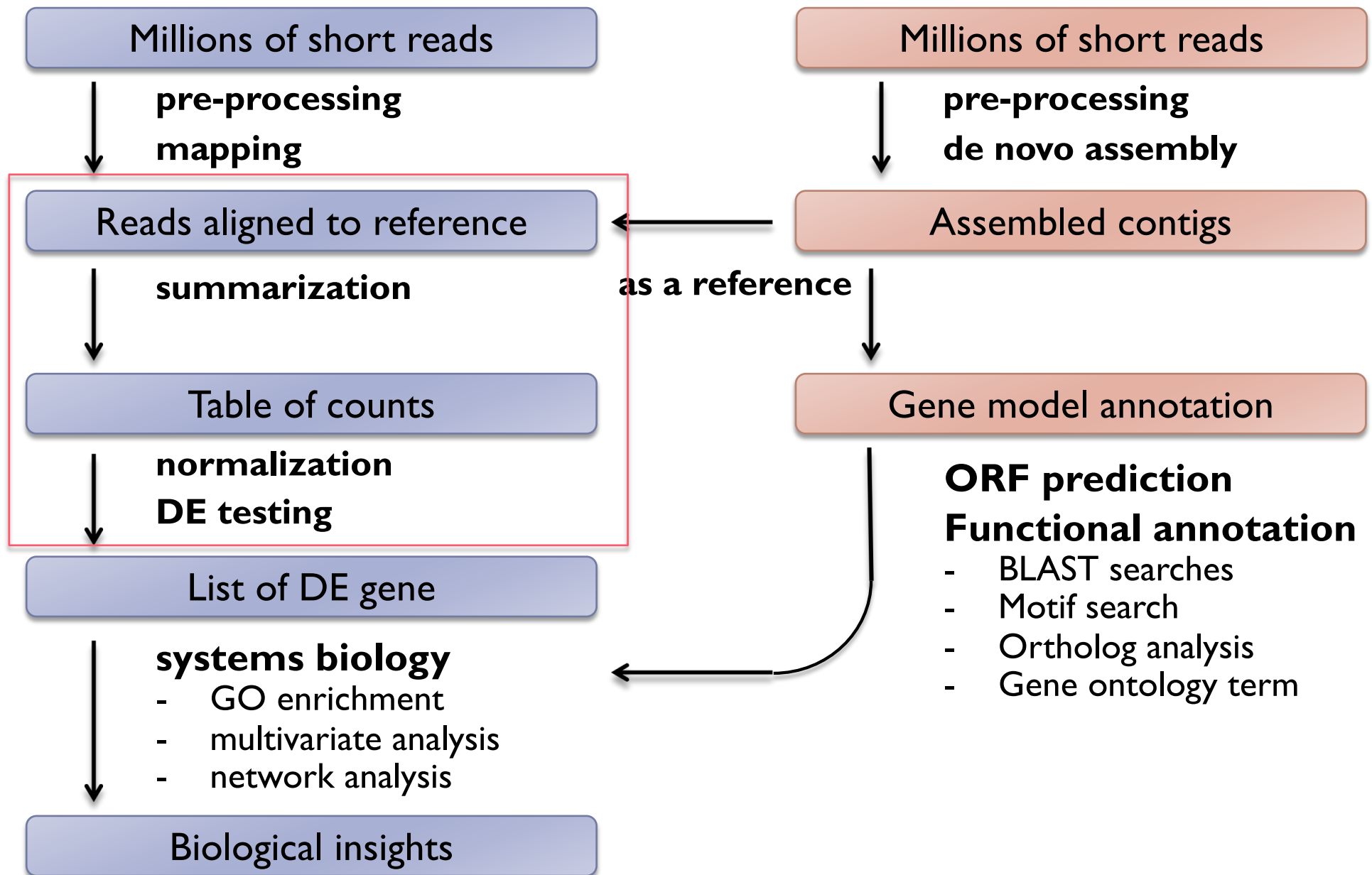
# Trinity example

- ▶ Input: Illumina short reads in FASTQ | FASTA format
- ▶ Output: assembled contigs in FASTA format

```
# Run Trinity
$ Trinity --seqType fq --left left_all.fq --right right_all.fq \
          --CPU 8 --max_memory 20G
```

(Trinity is supported on only Linux)

# RNA-seq analysis pipeline (*de novo* strategy)



# DEG analysis

- ▶ Follow transcript-based RNA-seq pipeline



# Clean up reference sequences

- ▶ **An issue: Inflation of the number of Trinity contigs is often observed.**
  - ▶ Trinity outputs splicing variants separately
  - ▶ Contaminations
  - ▶ Artifacts (bad contigs)
  - ▶ Incomplete contigs with very low expression.
- ▶ **Solution**
  - ▶ Filter out unwanted contigs.
  - ▶ Filter out very lowly expressed transcripts.
  - ▶ Cluster similar sequences.

# Remove redundancy in reference sequences

## ► Strategy and Tools

- Choose one representative transcript from each cluster based on Trinity component information. (longest or highest expression)

## ► Clustering

- CDHIT-EST (<http://weizhongli-lab.org/cd-hit/>)
- Corset (Davidson et al., 2014).
- RapClust (<https://github.com/COMBINE-lab/RapClust>)
- EvidentialGene  
(<http://arthropods.eugenes.org/EvidentialGene/trassembly.html>)

## ► Advantage of redundancy reduction

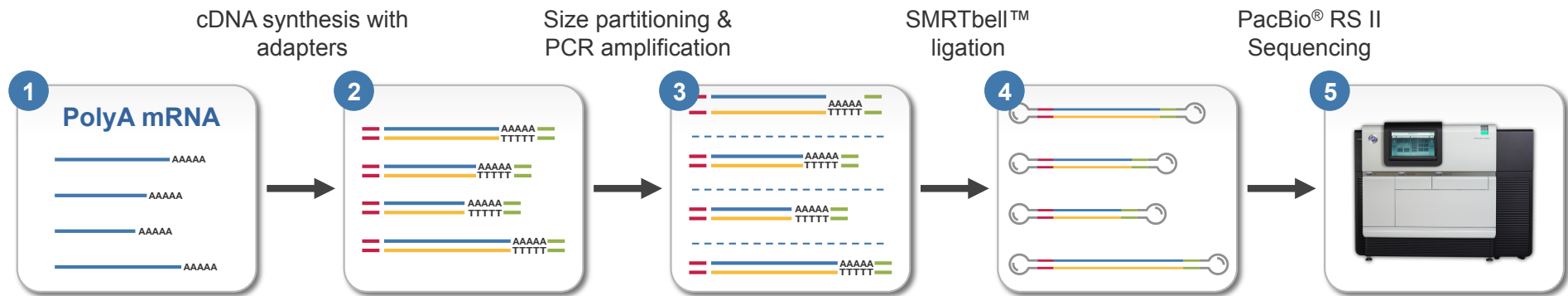
- Gene-oriented analysis => easier interpretation
- Better control of multiple comparison.

# Let's try Trinity assembly

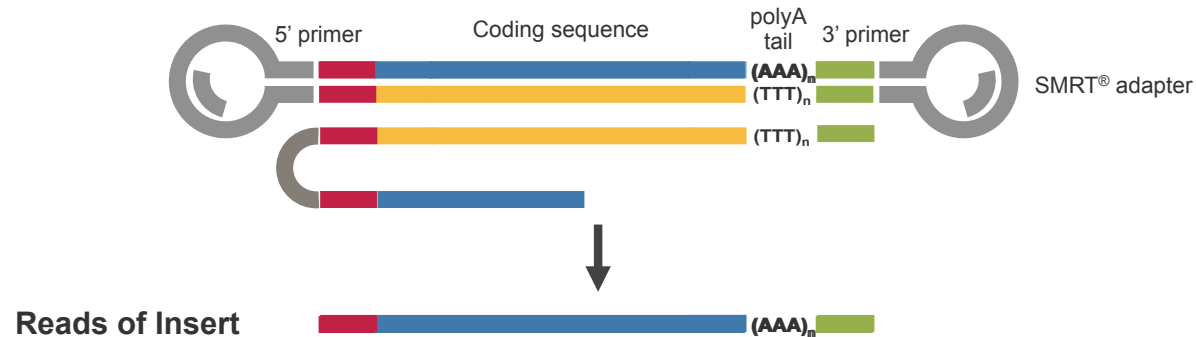
- ▶ ex9: de novo RNA-seq assembly using Trinity

# PacBio Iso-Seq for building a transcriptome catalogues

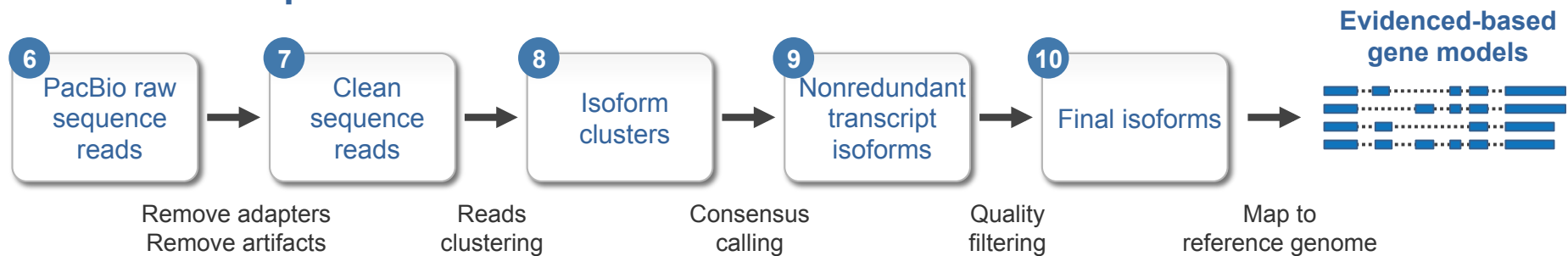
## Experimental Pipeline



## SampleNet: Iso-Seq Method with Clontech® cDNA Synthesis Kit



## Informatics Pipeline





# Others

- ▶ SuperTranscript (Davidson 2017)