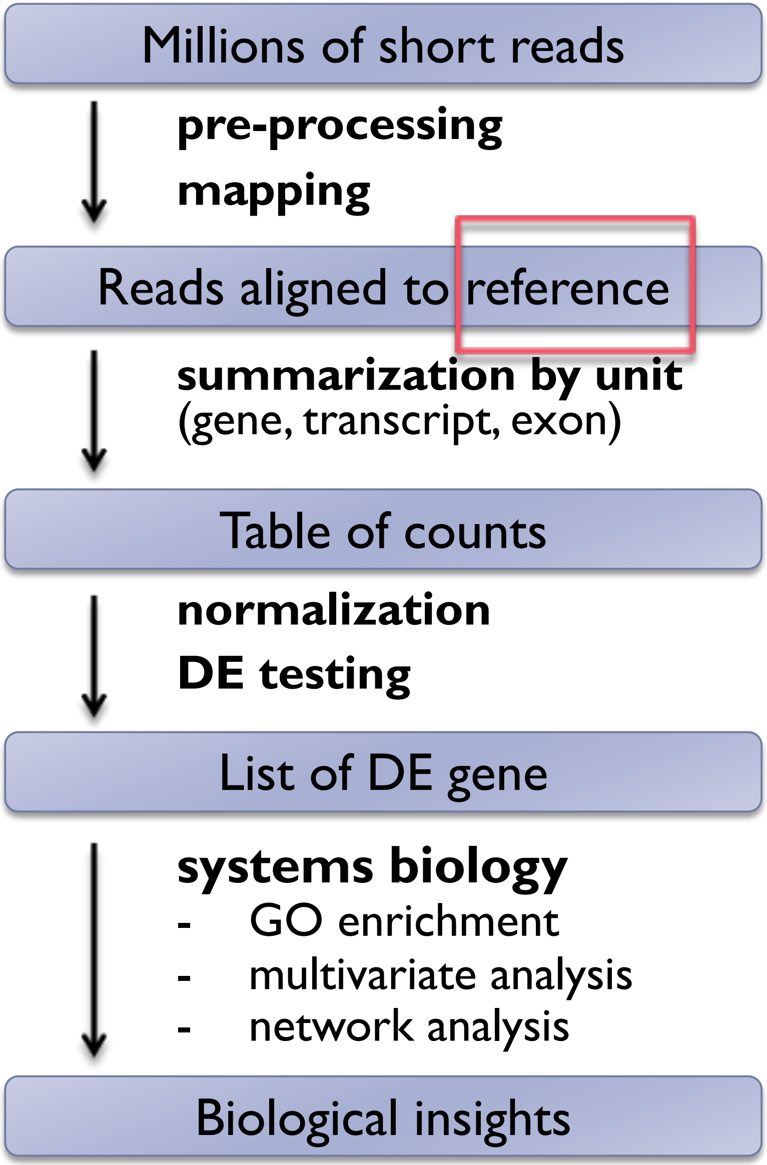
基礎生物学研究所　ゲノムインフォマティクス・トレーニングコース　2022年 3月

RNA-seq解析パイプライン：
de novo

Shuji Shigenobu
重信　秀治

基礎生物学研究所
生物機能解析センター

de novo RNA-seq



Millions of short reads

pre-processing

mapping

Reads aligned to reference

summarization by unit
(gene, transcript, exon)

Table of counts

normalization

DE testing

List of DE gene

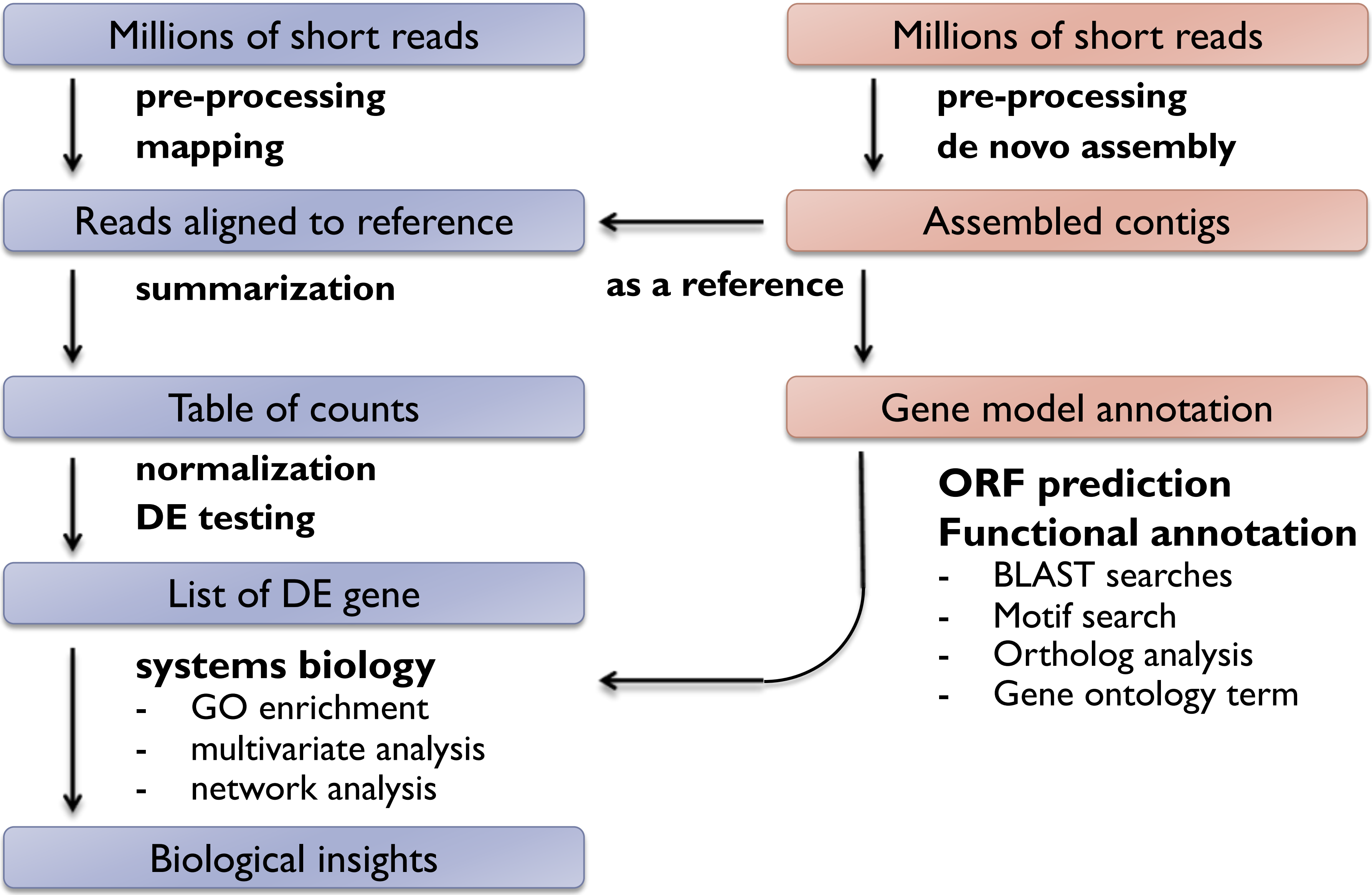
systems biology

- GO enrichment
- multivariate analysis
- network analysis

Biological insights

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

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



Millions of short reads

pre-processing

mapping

Reads aligned to reference

summarization

Table of counts

normalization

DE testing

List of DE gene

systems biology

- GO enrichment
- multivariate analysis
- network analysis

Biological insights

Millions of short reads

pre-processing

de novo assembly

Assembled contigs

as a reference

Gene model annotation

ORF prediction

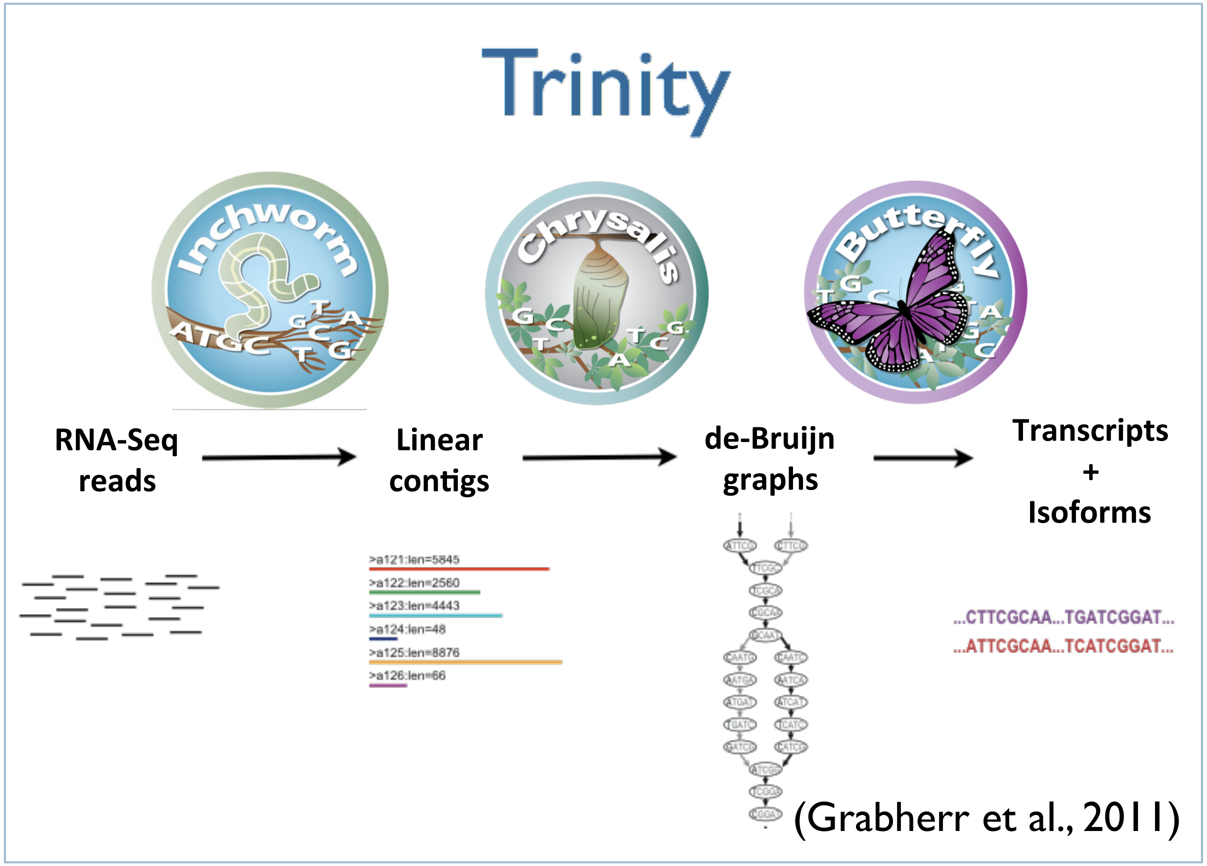
Functional annotation

- BLAST searches
- Motif search
- Ortholog analysis
- Gene ontology term

de novo assemblers of RNA-seq

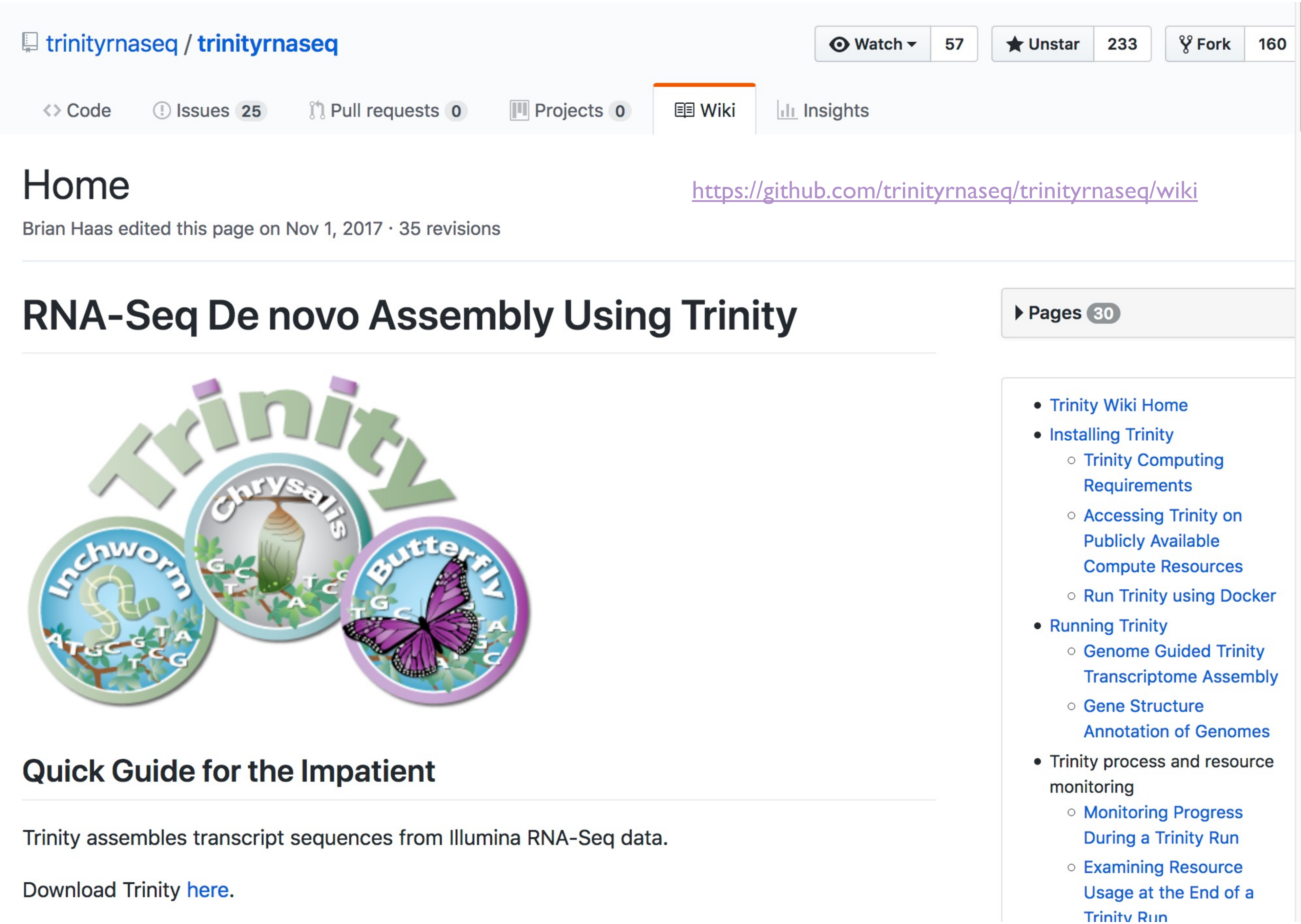
De novo assemblers use reads to assemble transcripts directly, which does not depend on a reference genome.

- Trinity
- Oases
- TransAbyss
- …



(Grabherr et al., 2011)

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



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

Home

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

RNA-Seq De novo Assembly Using Trinity

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)

Let’s try Trinity assembly

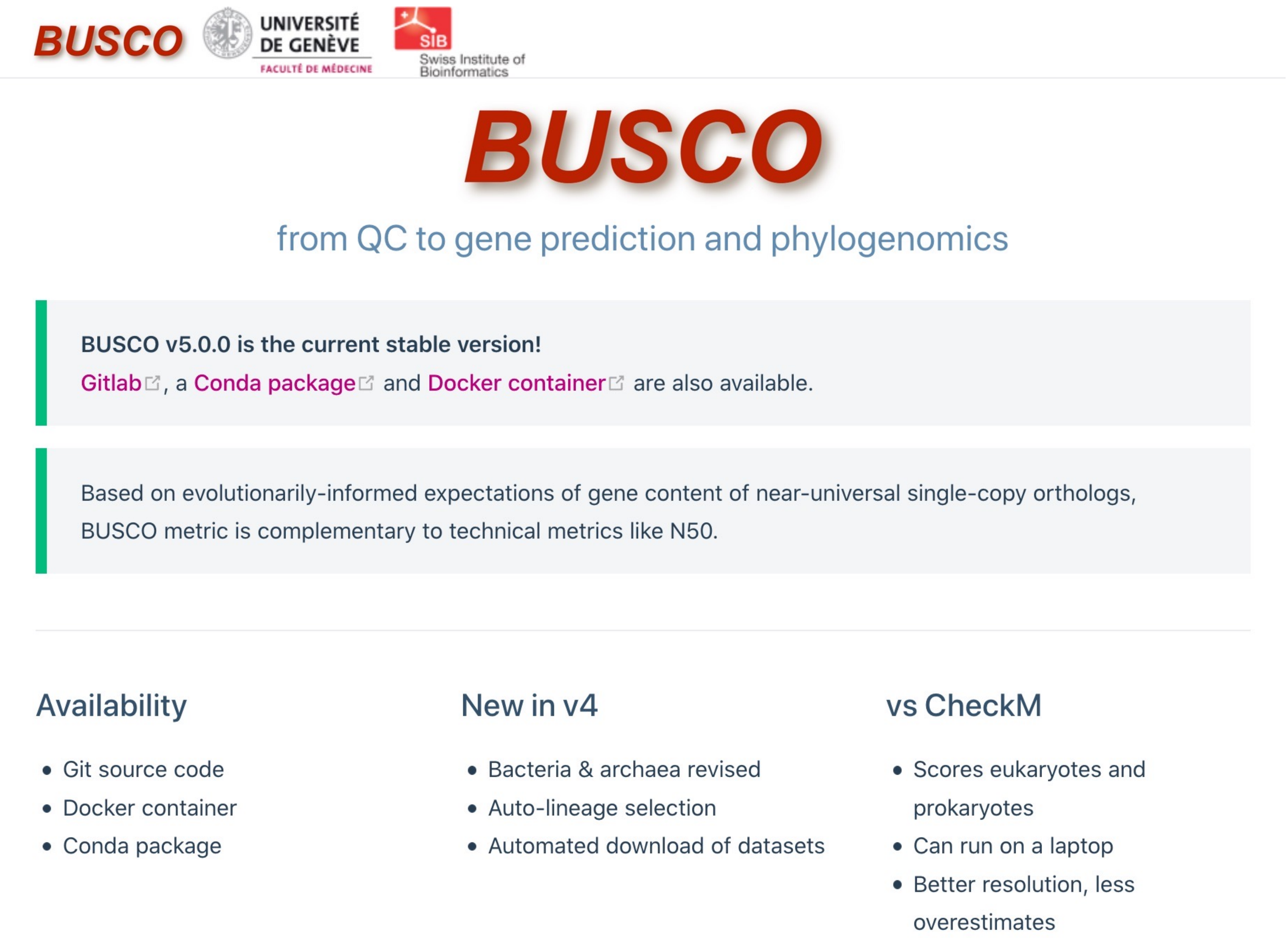
- ex701: *de novo* RNA-seq assembly using Trinity

Evaluate assembly

- Assembly stats
  - Number of contigs
  - Total length
  - mean, median, N50
- Coverage
  - BUSCO
  - Map back input reads
  - Map other RNAseq reads / known transcripts
- Contamination
  - BLAST (diamond) nr

BUSCO

https://busco.ezlab.org/



BUSCO

from QC to gene prediction and phylogenomics

BUSCO v5.0.0 is the current stable version!
Gitlab, a Conda package and Docker container are also available.

Based on evolutionarily-informed expectations of gene content of near-universal single-copy orthologs, BUSCO metric is complementary to technical metrics like N50.

Availability

- Git source code
- Docker container
- Conda package

New in v4

- Bacteria & archaea revised
- Auto-lineage selection
- Automated download of datasets

vs CheckM

- Scores eukaryotes and prokaryotes
- Can run on a laptop
- Better resolution, less overestimates

BUSCO

BUSCO provides a quantitative assessment of the completeness in terms of expected gene content of a genome assembly or transcriptome by using universally conserved one-copy gene set. The results are simplified into categories of Complete and single-copy, Complete and duplicated, Fragmented, or Missing.

```
# Run BUSCO
$ busco –m transcriptome contigs.fa –o OUTPUT –l lineage
```

```
# example of output
    (Insecta)
    C:94.5%[S:88.5%,D:6.0%],F:1.1%,M:4.4%,n:978

    925 Complete BUSCOs (C)
    866 Complete and single-copy BUSCOs (S)
    59  Complete and duplicated BUSCOs (D)
    11  Fragmented BUSCOs (F)
    42  Missing BUSCOs (M)
    978 Total BUSCO groups searched
```

練習：ex702

Advanced

Clean up reference sequences

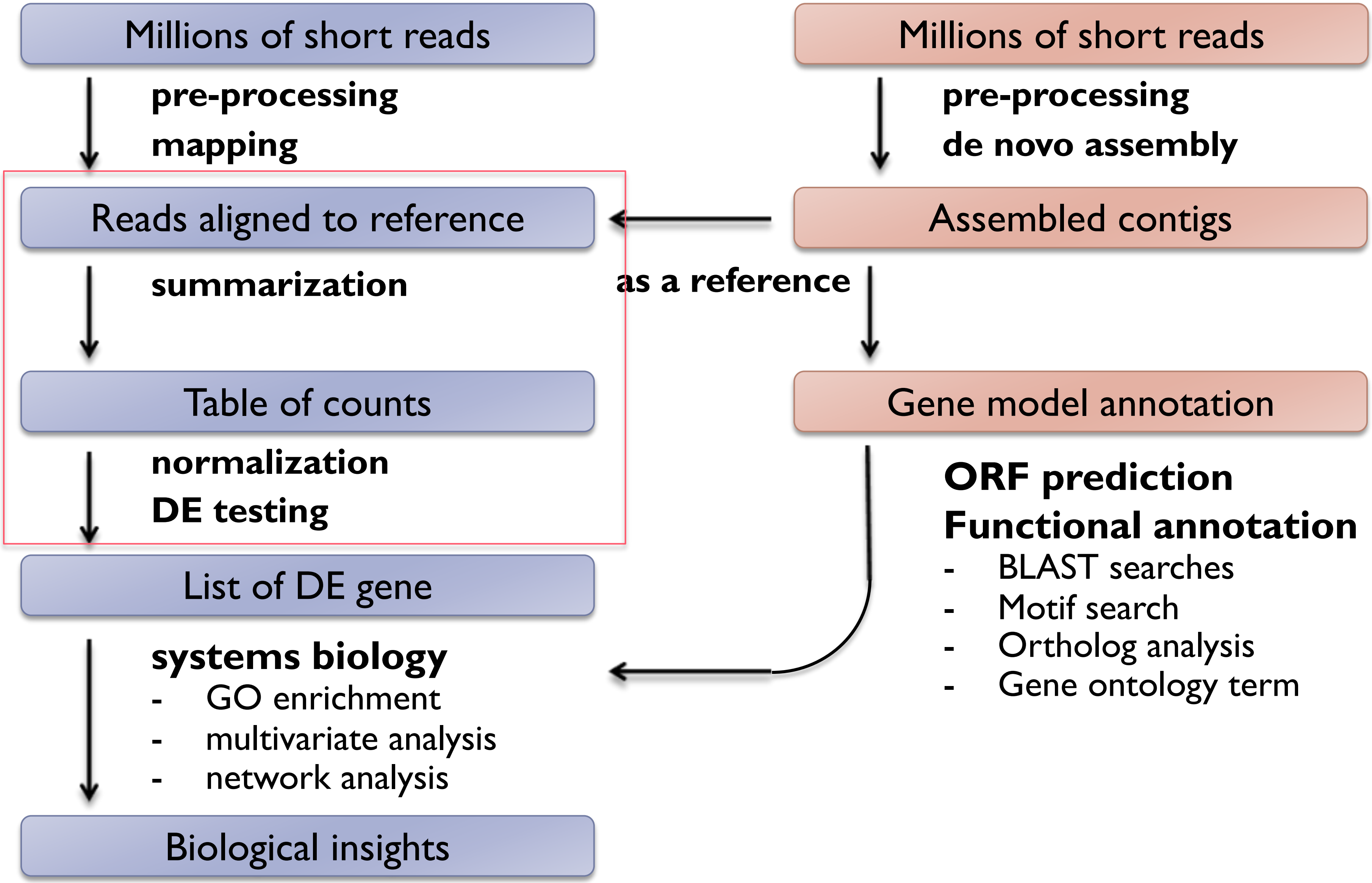
- An issue: Inflation of the number of Trinity contigs is often observed.
  - Trinity outputs splicing variants separately
  - Contaminations
  - Artifacts (bad contigs)
  - Incomplete contigs with very low expression.
- Solution
  - Filter out unwanted contigs.
  - Filter out very lowly expressed transcripts.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

Advanced

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.

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



Millions of short reads

pre-processing

mapping

Reads aligned to reference

summarization

Table of counts

normalization

DE testing

List of DE gene

systems biology

- GO enrichment
- multivariate analysis
- network analysis

Biological insights

Millions of short reads

pre-processing

de novo assembly

Assembled contigs

as a reference

Gene model annotation

ORF prediction

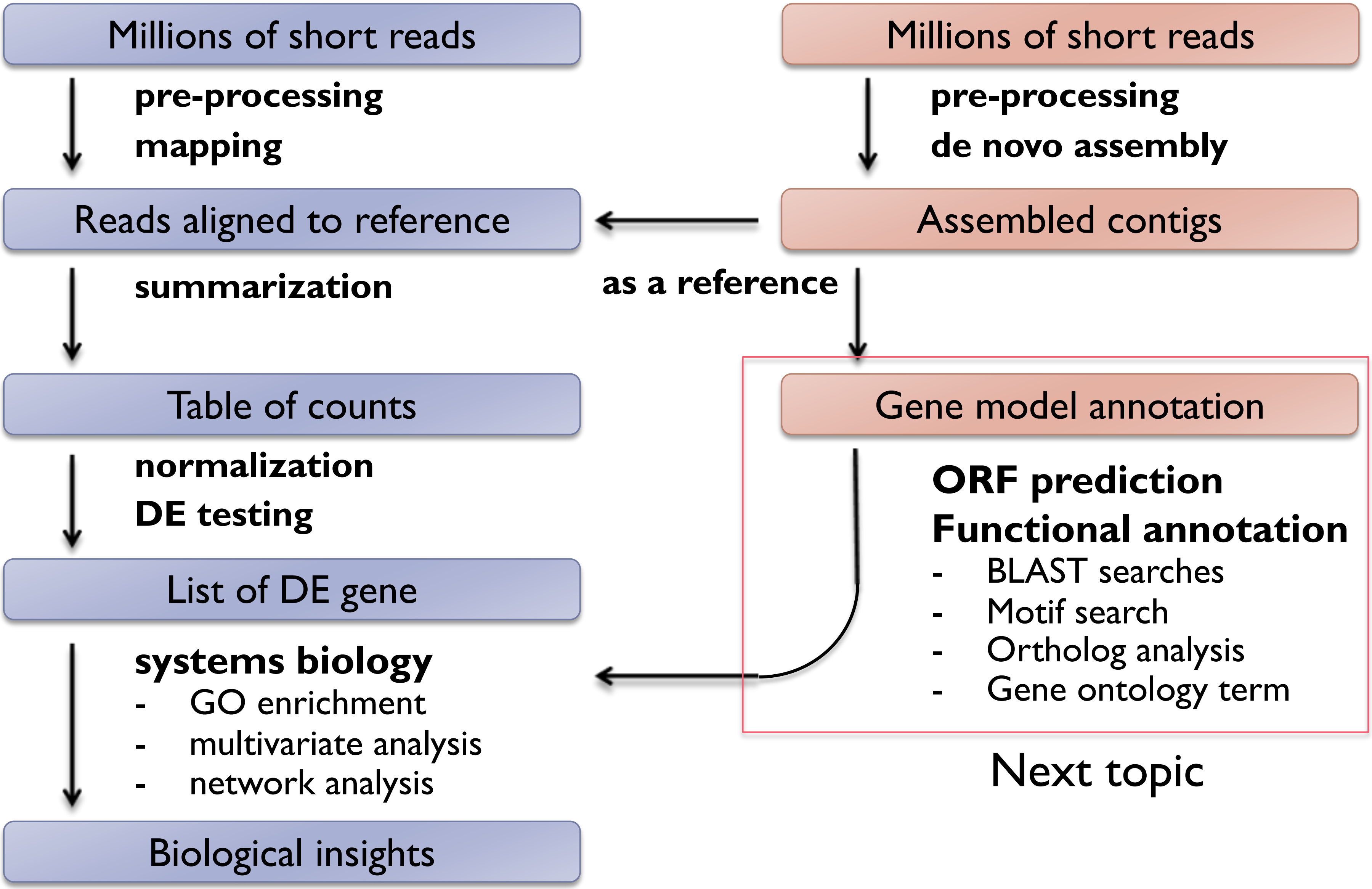
Functional annotation

- BLAST searches
- Motif search
- Ortholog analysis
- Gene ontology term

DEG analysis

- Follow transcript-based RNA-seq pipeline

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



Millions of short reads

pre-processing

mapping

Reads aligned to reference

summarization

Table of counts

normalization

DE testing

List of DE gene

systems biology

- GO enrichment
- multivariate analysis
- network analysis

Biological insights

Millions of short reads

pre-processing

de novo assembly

Assembled contigs

as a reference

Gene model annotation

ORF prediction

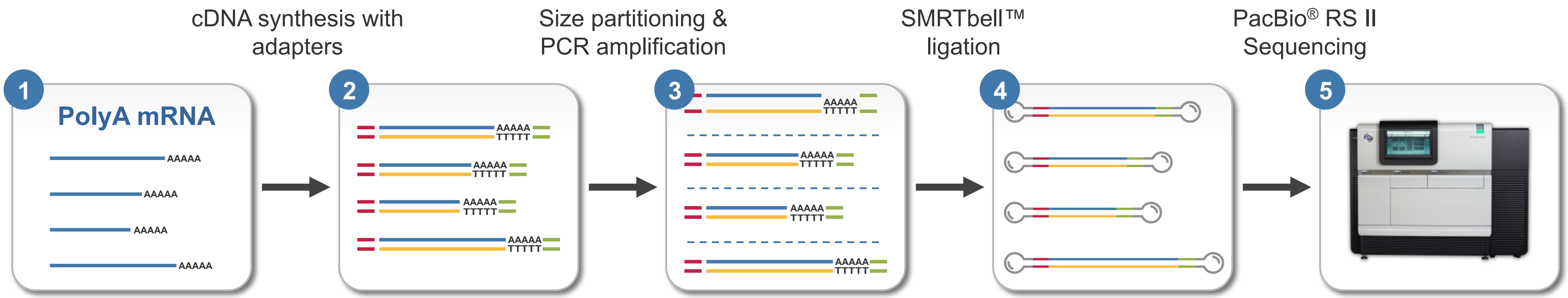
Functional annotation

- BLAST searches
- Motif search
- Ortholog analysis
- Gene ontology term

Next topic

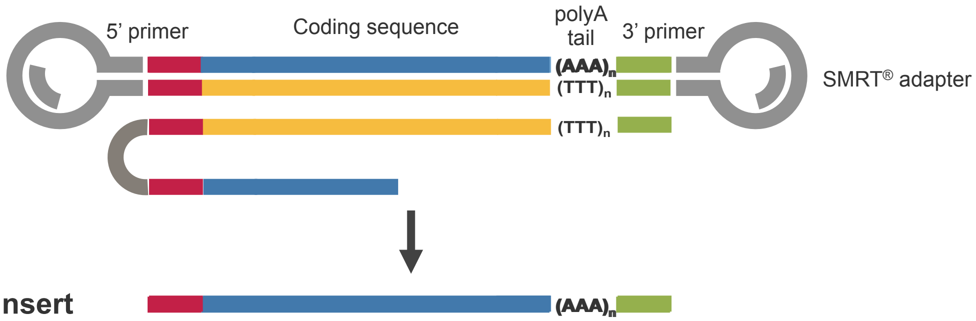
PacBio Iso-Seq for building a transcriptome catalogues

Experimental Pipeline



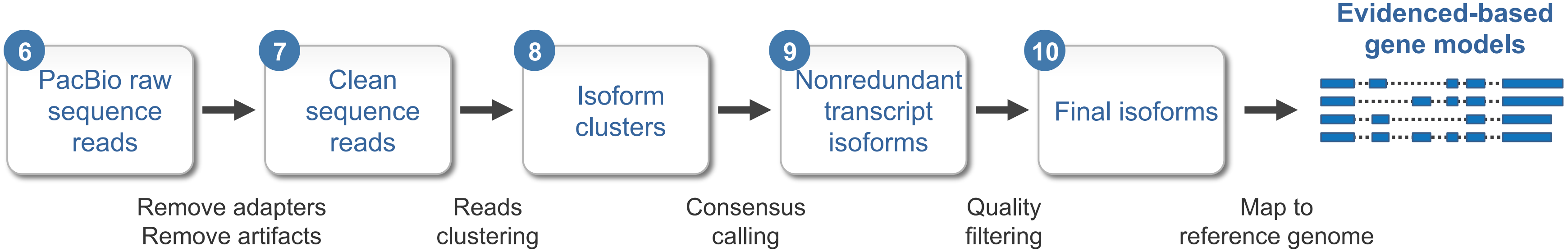
1. PolyA mRNA
2. cDNA synthesis with adapters
3. Size partitioning & PCR amplification
4. SMRTbell™ ligation
5. PacBio® RS II Sequencing

SampleNet: Iso-Seq Method with Clonetech® cDNA Synthesis Kit



Reads of Insert

Informatics Pipeline



6. PacBio raw sequence reads
7. Clean sequence reads
8. Isoform clusters
9. Nonredundant transcript isoforms
10. Final isoforms

Evidenced-based gene models

Remove adapters
Remove artifacts

Reads clustering

Consensus calling

Quality filtering

Map to reference genome

N50

- N50
- 2000 3000 100 6000 5000
- len.sorted <- rev(sort(len))
- N50 <- len.sorted[cumsum(len.sorted) >= sum(len.sorted)*0.5][1]

Others

- SuperTranscript (Davidson 2017)