

# RNA-seq解析パイプライン： *de novo*

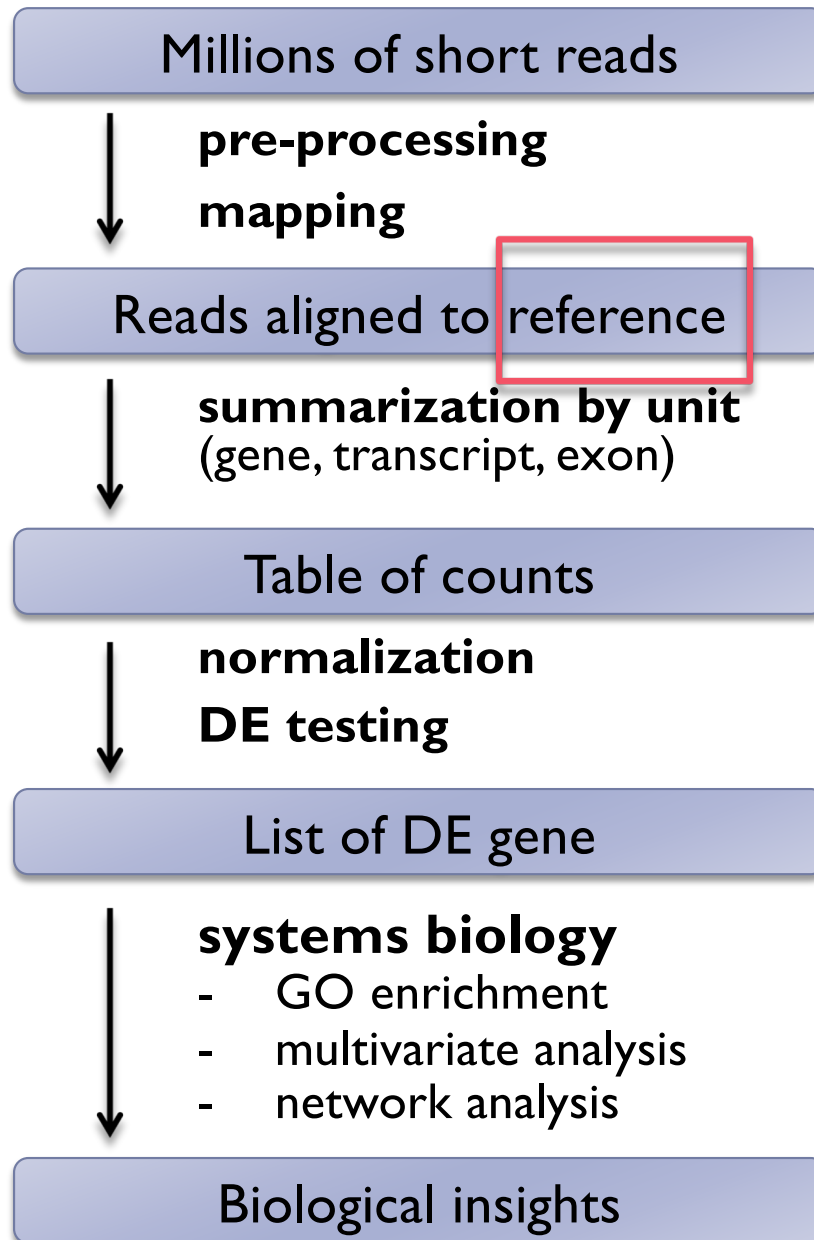
Shuji Shigenobu

重信 秀治

基礎生物学研究所  
生物機能解析センター

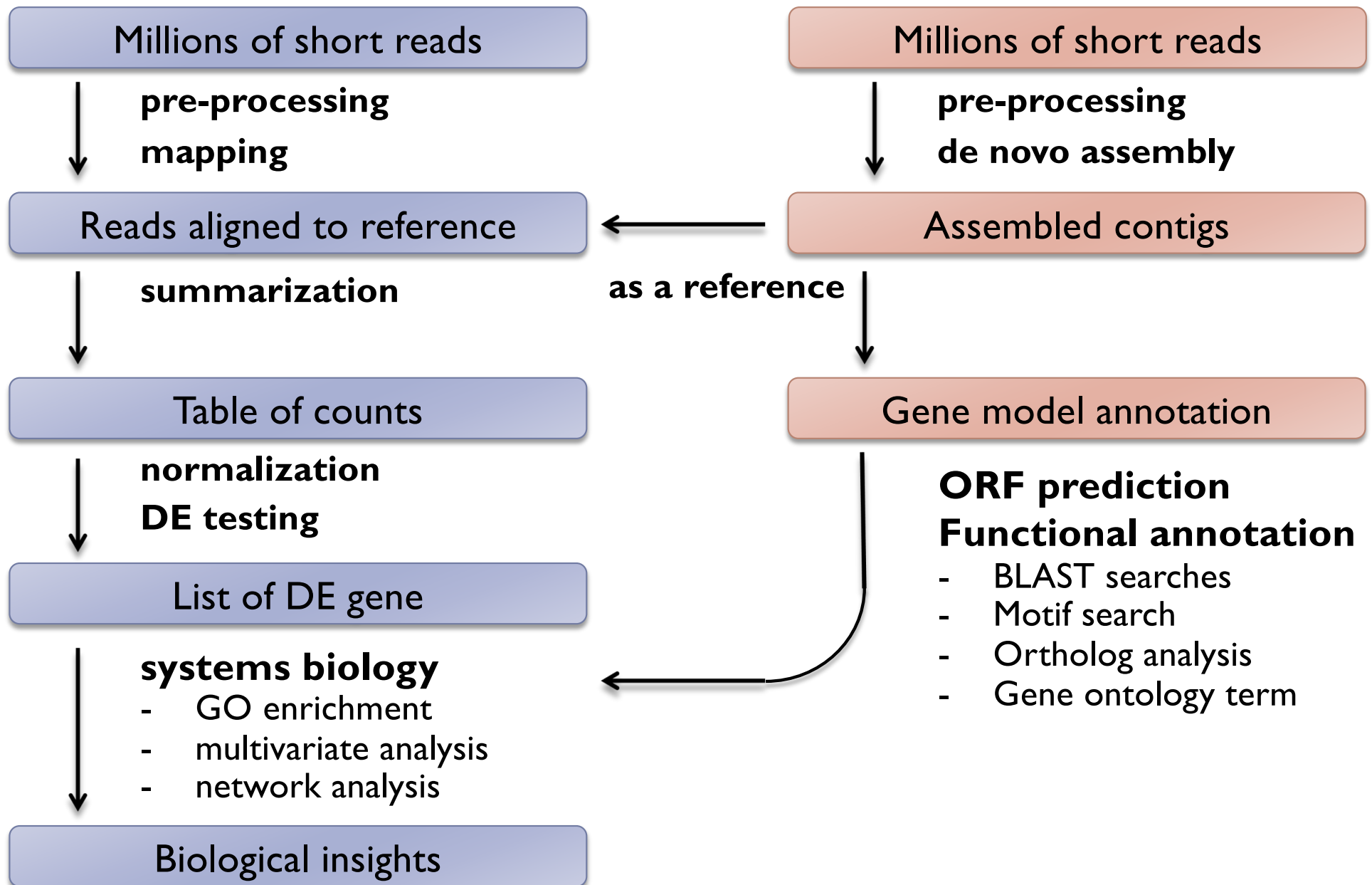


# *de novo* RNA-seq



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

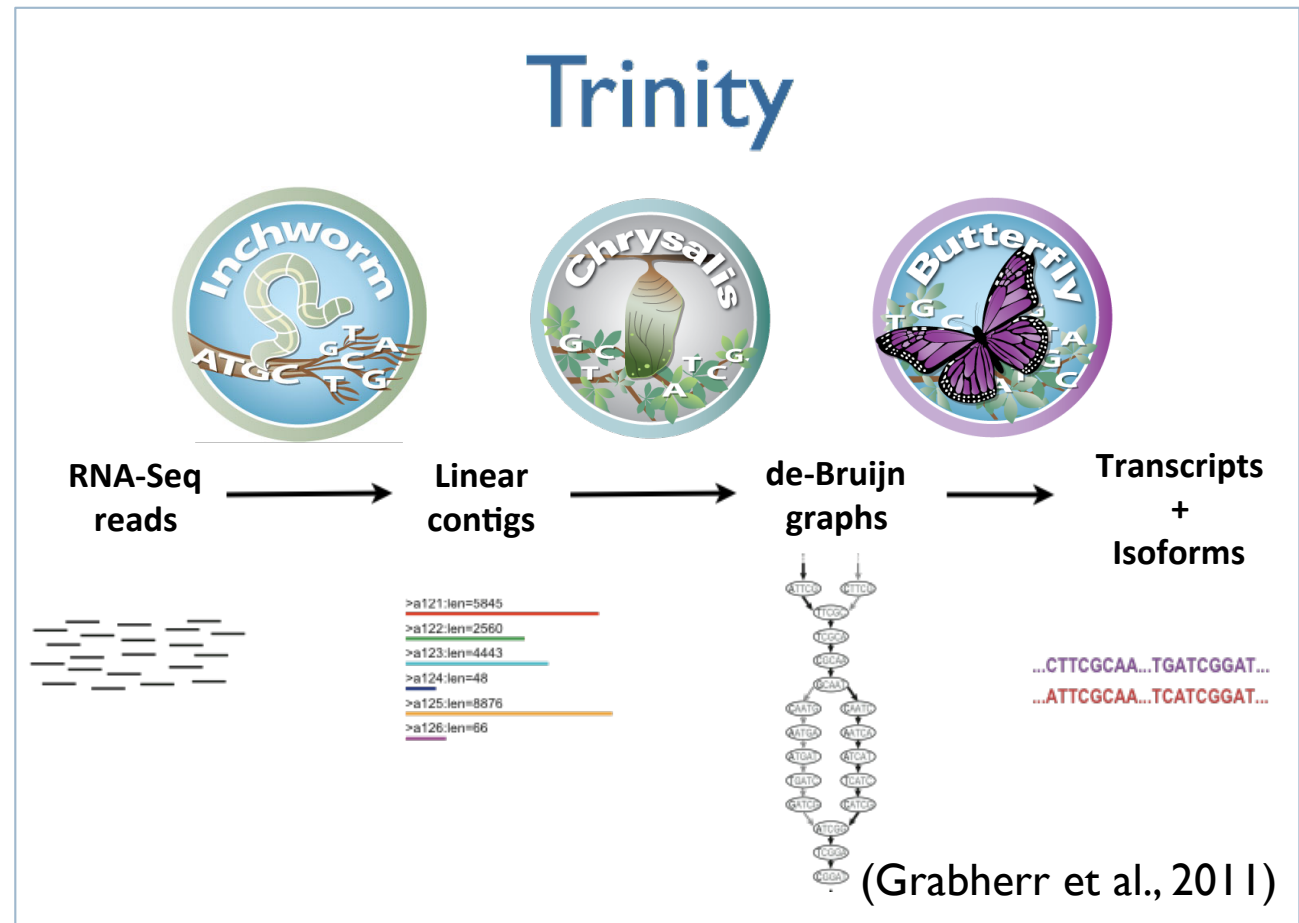
# RNA-seq analysis pipeline (*de novo* strategy)



# *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)

# 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





UNIVERSITÉ  
DE GENÈVE  
FACULTÉ DE MÉDECINE



Swiss Institute of  
Bioinformatics

# 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
```

# 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

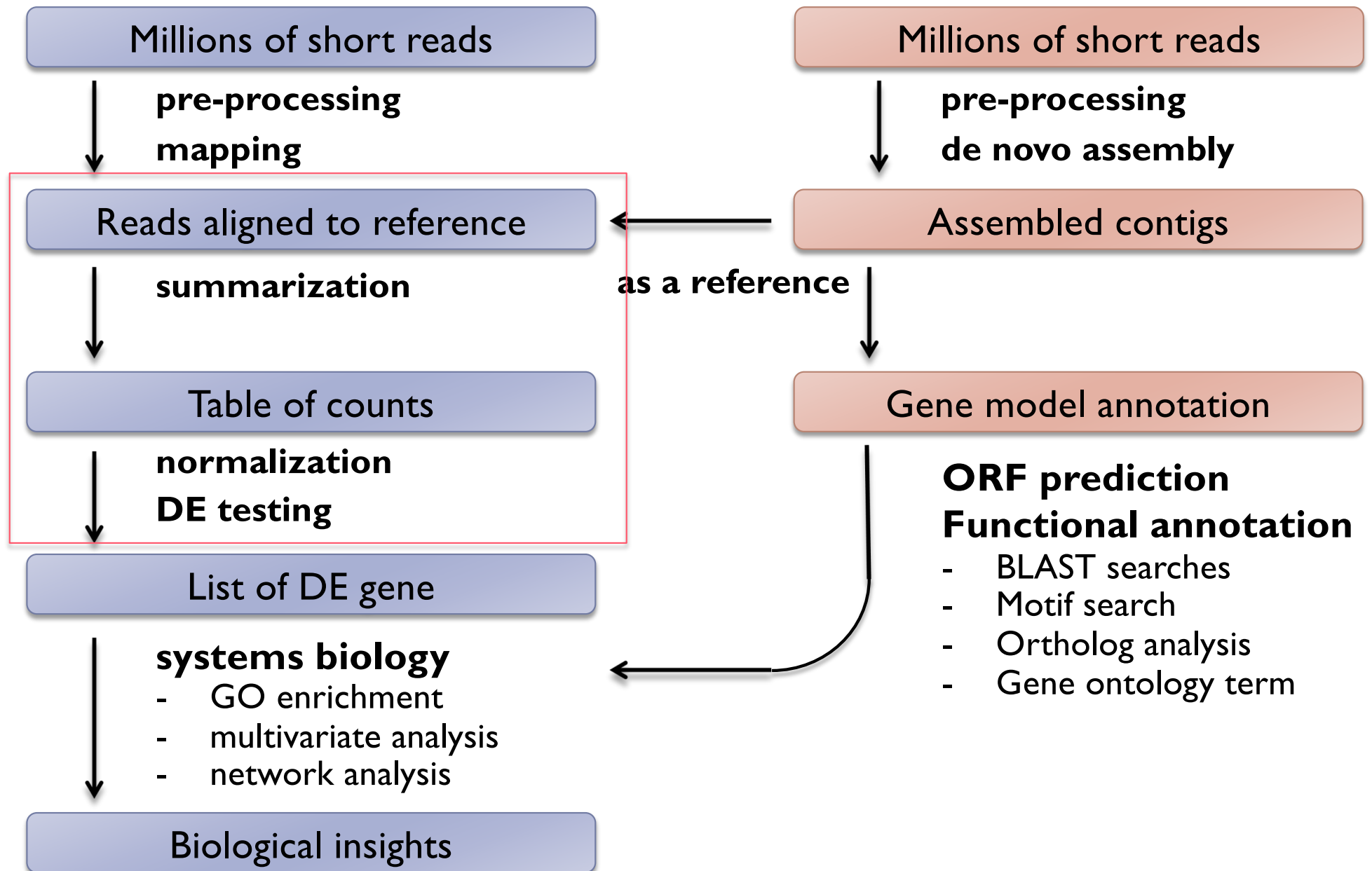
## ► Strategy and Tools

- 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.

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



# DEG analysis

- ▶ Follow transcript-based RNA-seq pipeline