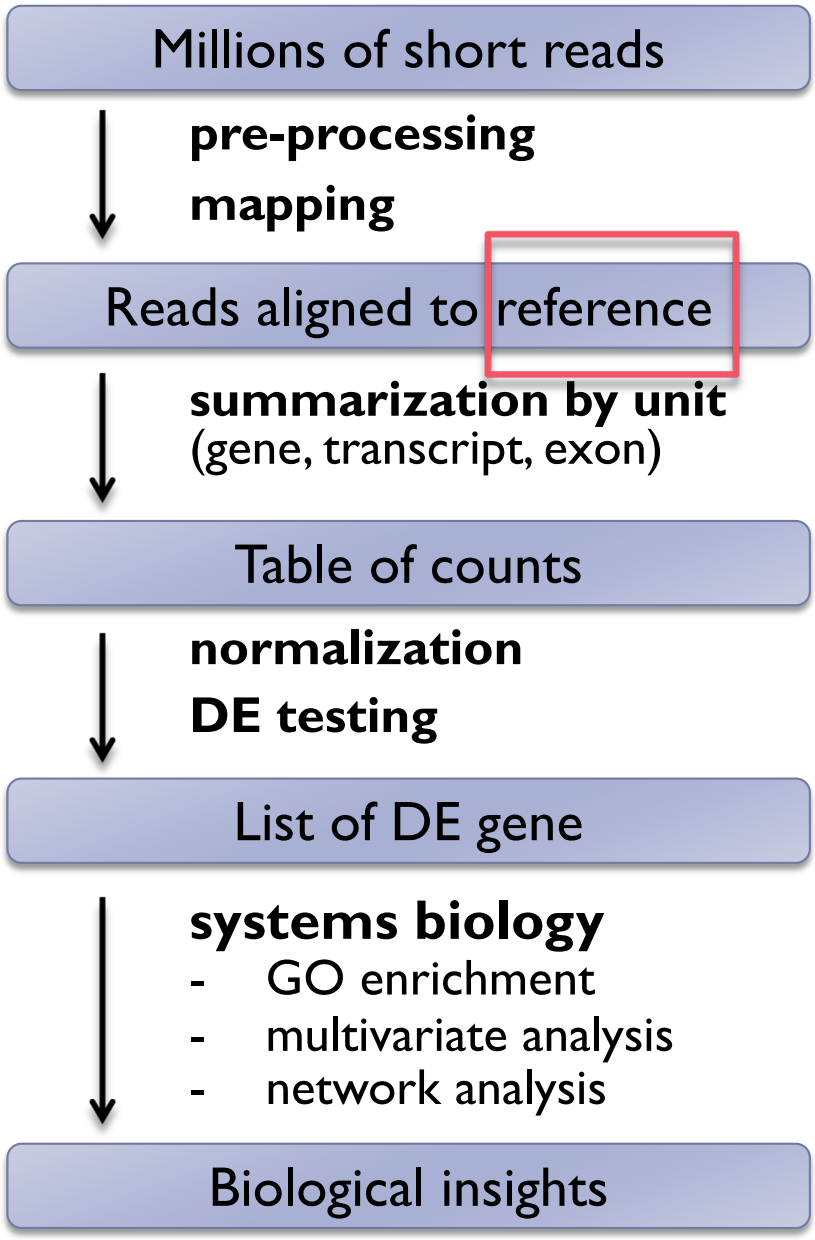
RNA-seq解析パイプライン：*de novo*

Shuji Shigenobu
重信　秀治

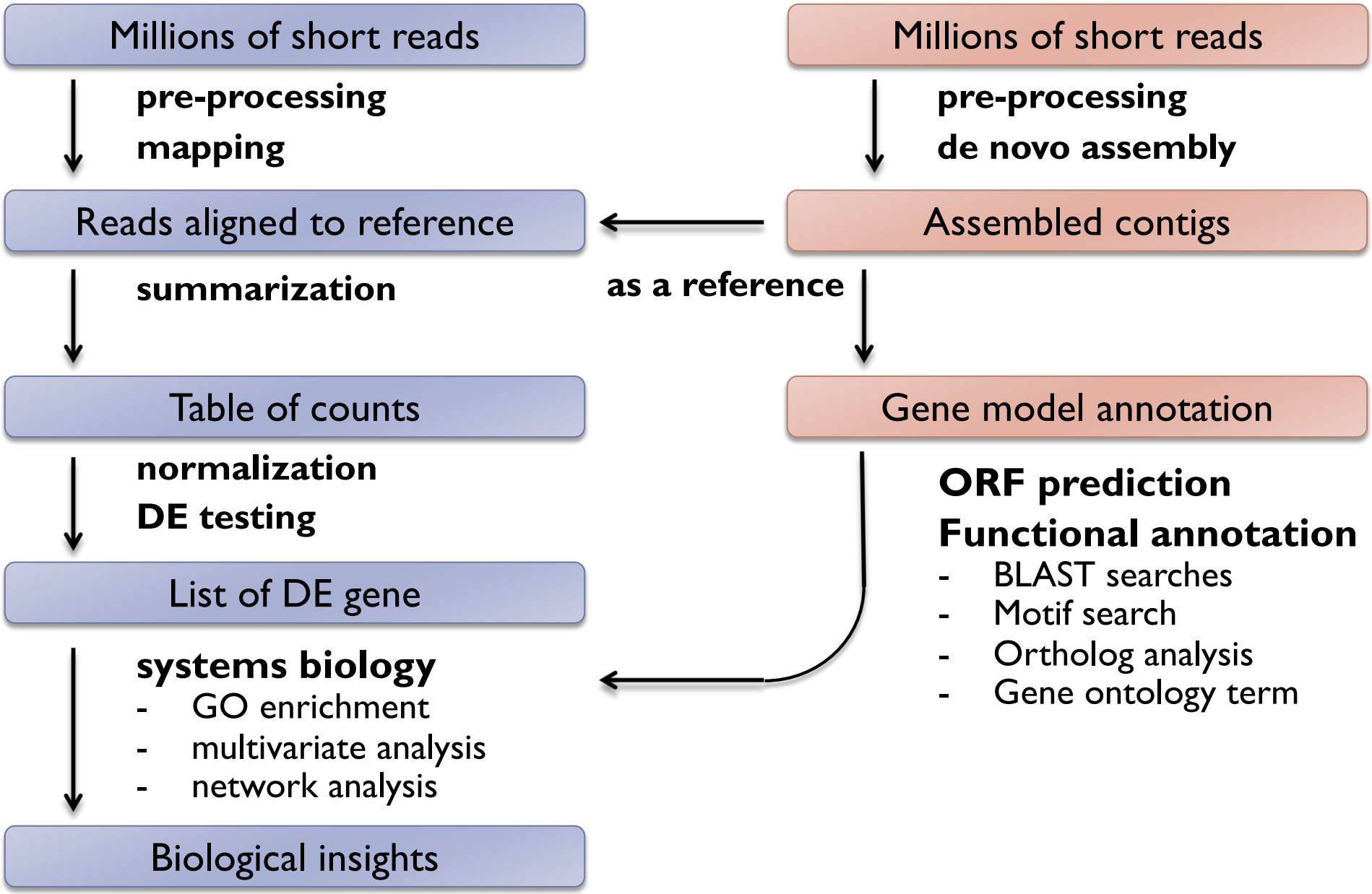
基礎生物学研究所
生物機能解析センター

de novo RNA-seq



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

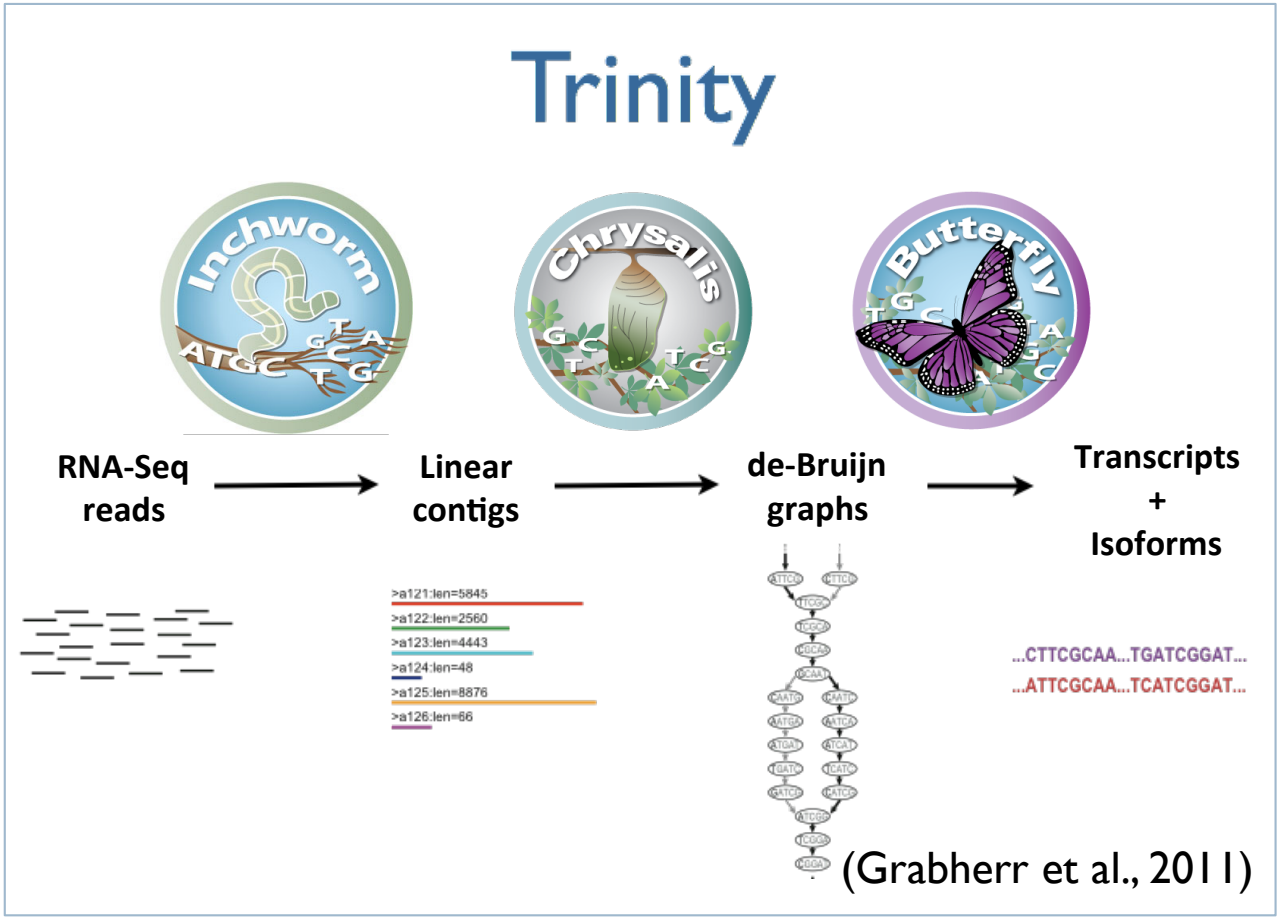
RNA-seq analysis pipeline (***de novo*** strategy)



de novo assemblers of RNA-seq

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

- Trinity
- Oases
- TransAbyss
- …



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

Home

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

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 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

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

Evaluate assembly

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

Advanced

Clean up reference sequences

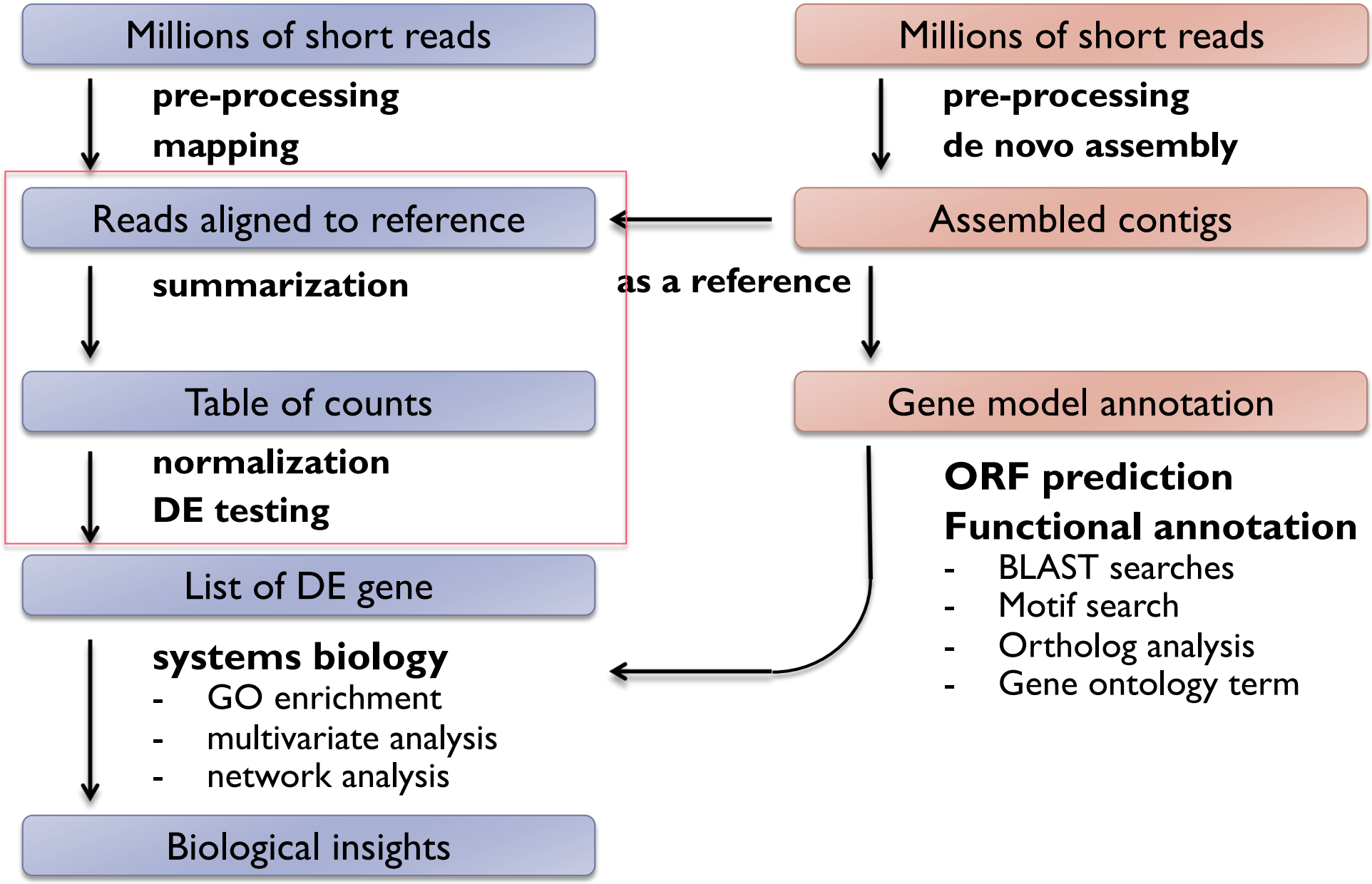
- An issue: Inflation of the number of Trinity contigs is often observed.
  - Trinity outputs splicing variants separately
  - Contaminations
  - Artifacts (bad contigs)
  - Incomplete contigs with very low expression.
- Solution
  - Filter out unwanted contigs.
  - Filter out very lowly expressed transcripts.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



DEG analysis

- Follow transcript-based RNA-seq pipeline

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