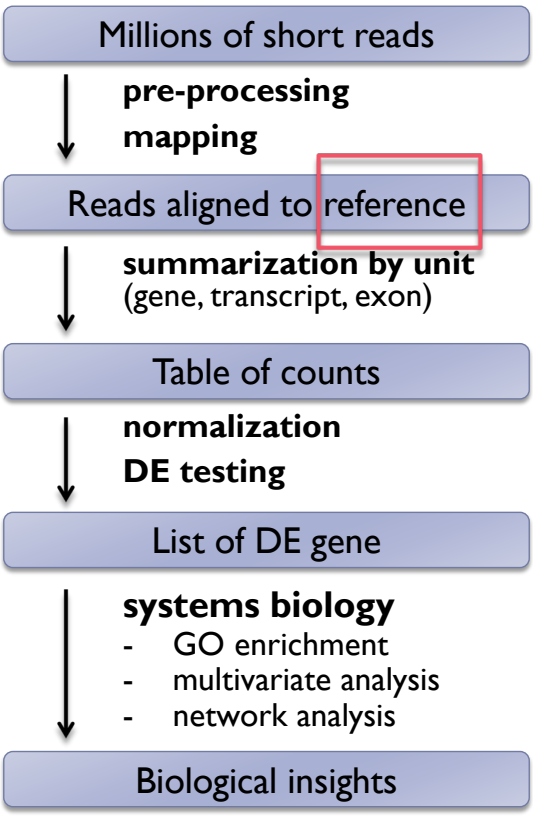
RNA-seq解析パイプライン：*de novo*

Shuji Shigenobu
重信　秀治

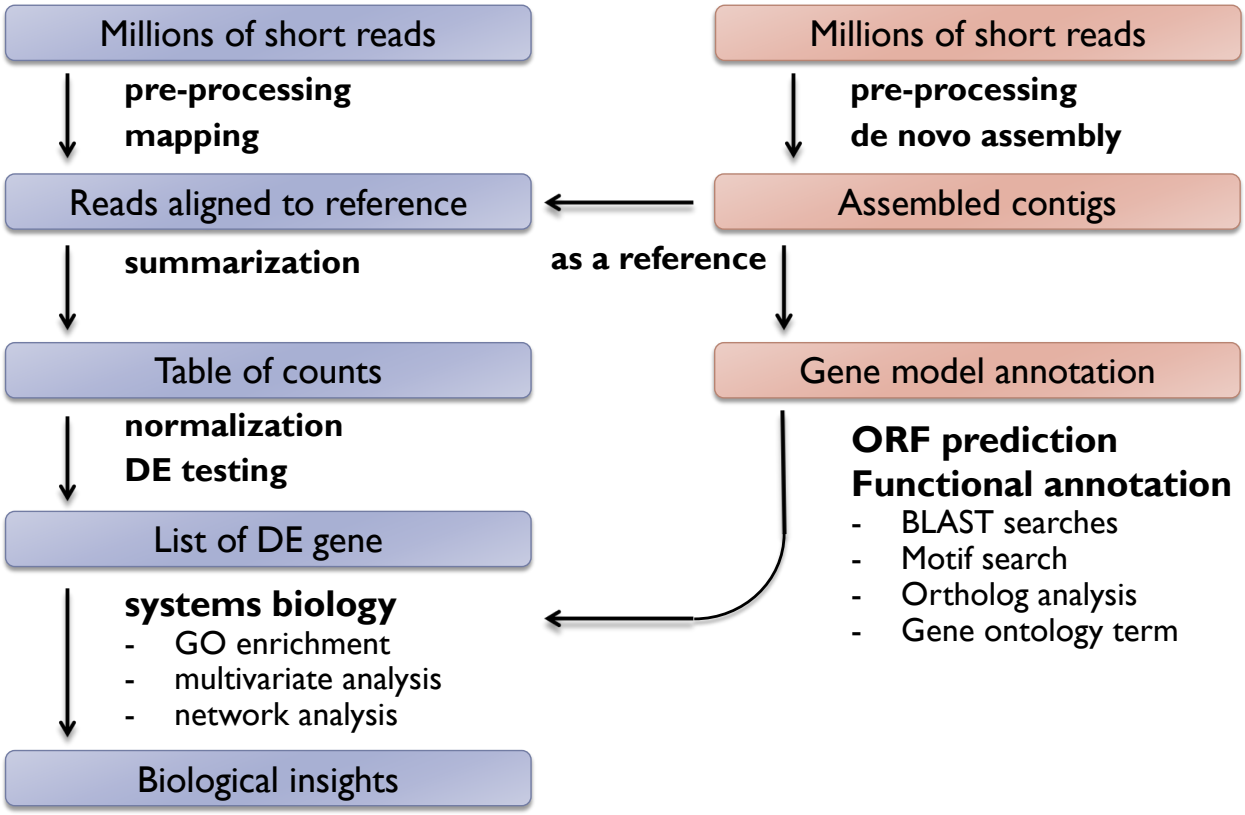
基礎生物学研究所
生物機能解析センター

de novo RNA-seq



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

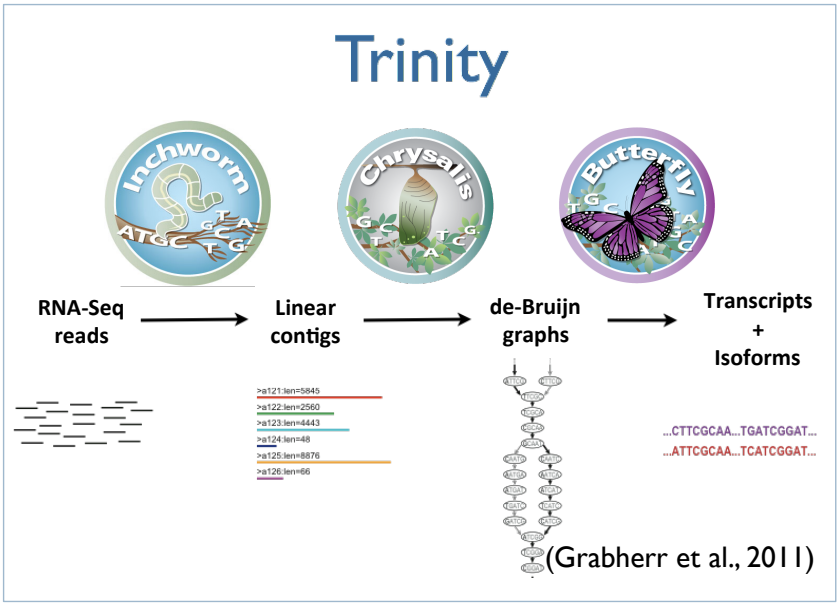
RNA-seq analysis pipeline (***de novo*** strategy)



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
- EBARDenovo
- …



http://trinityrnaseq.sourceforge.net/

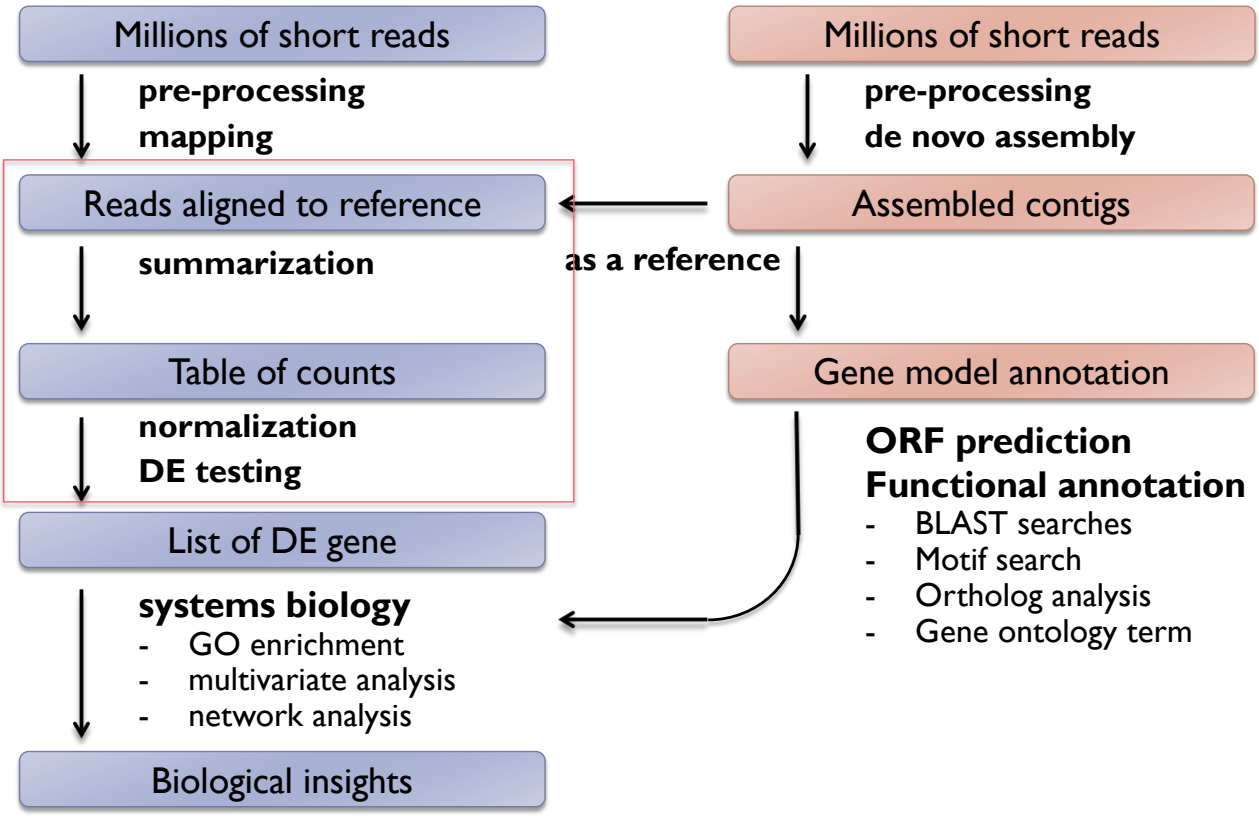
Trinity example

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

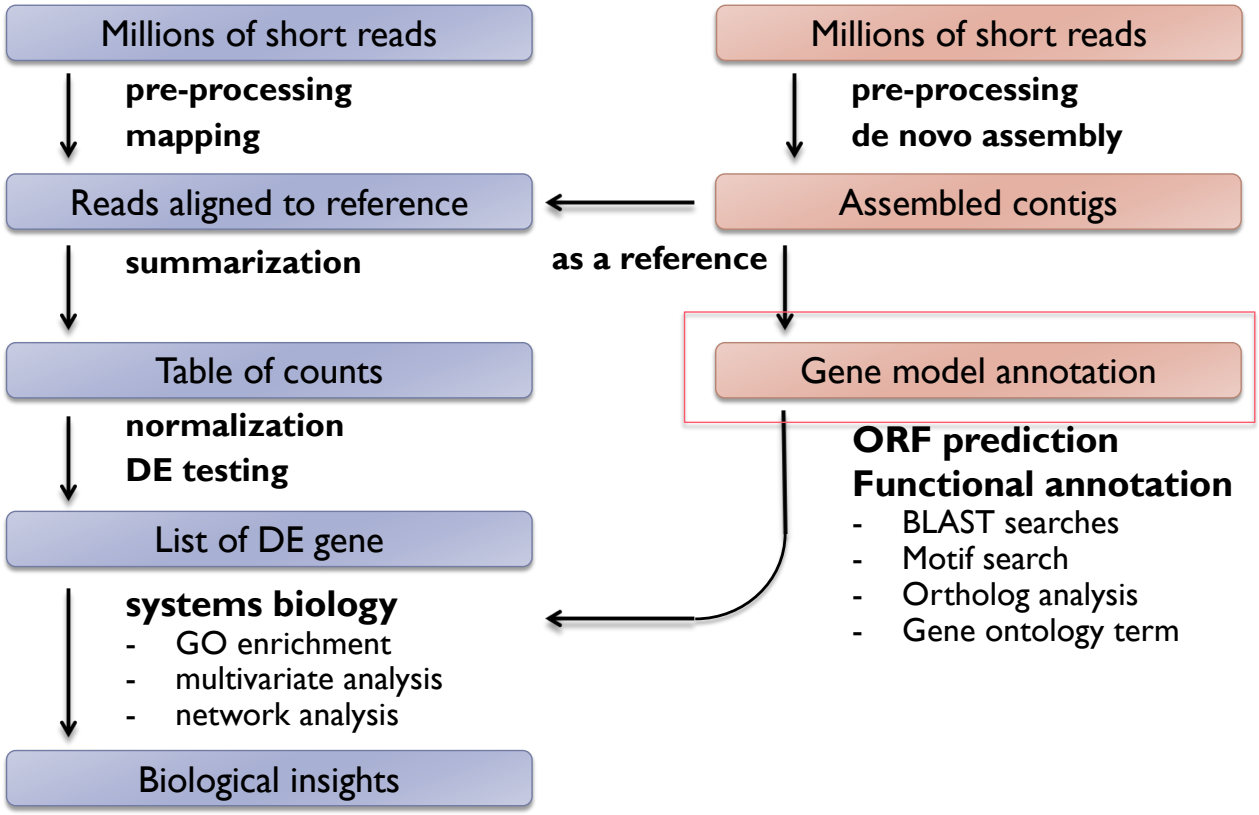
```
# Run Trinity
$ Trinity --seqType fq --left left_all.fq --right right_all.fq \
          --CPU 8 --max_memory 20G
```

(Trinity is supported on only Linux)

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



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



ORF prediction

- Special consideration in ORF prediction after de novo RNA-seq assembly
  - Sometimes partial: Start Met or terminal codon may be missing.
  - Ideally one ORF is present per contig, but erroneously joined contigs may include multiple ORFs.
  - Possible frame shifts.
    - Frame shifts do not occur so often in Illumina, while it happens very frequently in 454 and IonProton.
- Recommended software: TransDecoder

Functional Annotation of Predicted ORFs

- BLAST
  - NCBI NR (or UniProt)
  - species of interest (model organisms, close relatives etc)
  - specific DB (SwissProt, rRNA DB, CEGMA etc)
  - self (assembly v.s. assembly)
- Motif search
  - Pfam, SignalP etc.
- Ortholog analysis
  - vs model organism
  - ortholog database (OrthoDB, eggNOG, OrthoMCL etc)
  - close relatives
- Gene Ontology term assignment

Quick annotation by BLASTX

- Query: assembled contigs

  (nucleotide sequences in multi-fasta format)
- DB: Protein sequences of a model organism

Format DB

```
$ makeblastdb –in protein.fa -dbtype prot
```

Search

```
$ blastx -query trinity_contigs –db protein.fa \
  -num_threads 8 -evalue 1.0e-8 –outfmt 0 > blastxout.txt
```

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

「BLAST自由自在〜配列解析の極意をマスターする」

日時：2016年12月1日（木）10：30〜17：30

場所：基礎生物学研究所　（愛知県岡崎市）

受講申込み終了

2018年は秋に開催予定

講師：

重信 秀治 （基礎生物学研究所 生物機能解析センター 特任准教授）

内山 郁夫 （基礎生物学研究所 生物機能解析センター 助教）

重信 秀治　特任准教授

内山 郁夫　助教

Protein motif search using InterProScan

- Query: Translated ORF sequences
- Software: InterProScan
  - https://github.com/ebi-pf-team/interproscan/wiki

Search

```
$ interproscan.sh  -I proteins.fasta -f XML,TSV --goterms
--pathways
```

Assign Gene Ontology terms

- Tools
  - InterProScan
  - BLAST2GO
  - Transfer model organisms GO terms based on orthology.