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　2019年 3月

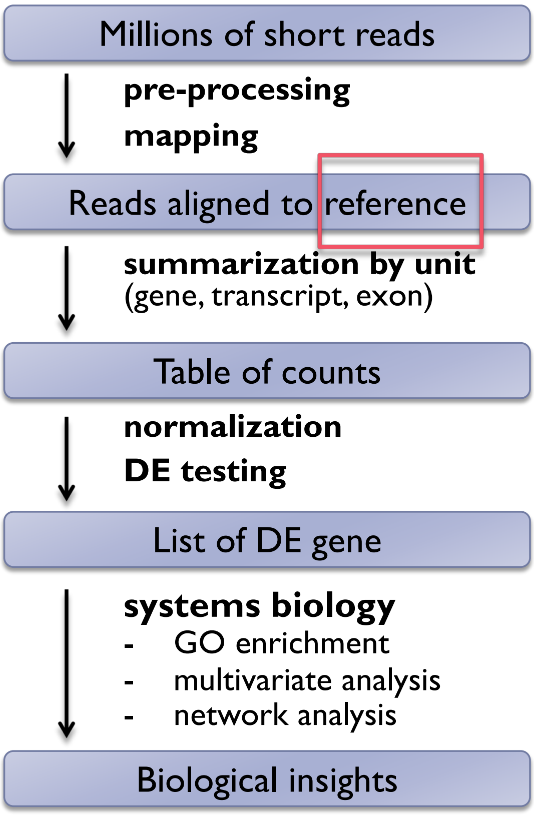
RNA-seq解析パイプライン： *de novo*

Shuji Shigenobu
重信　秀治

基礎生物学研究所
生物機能解析センター

NIBB

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

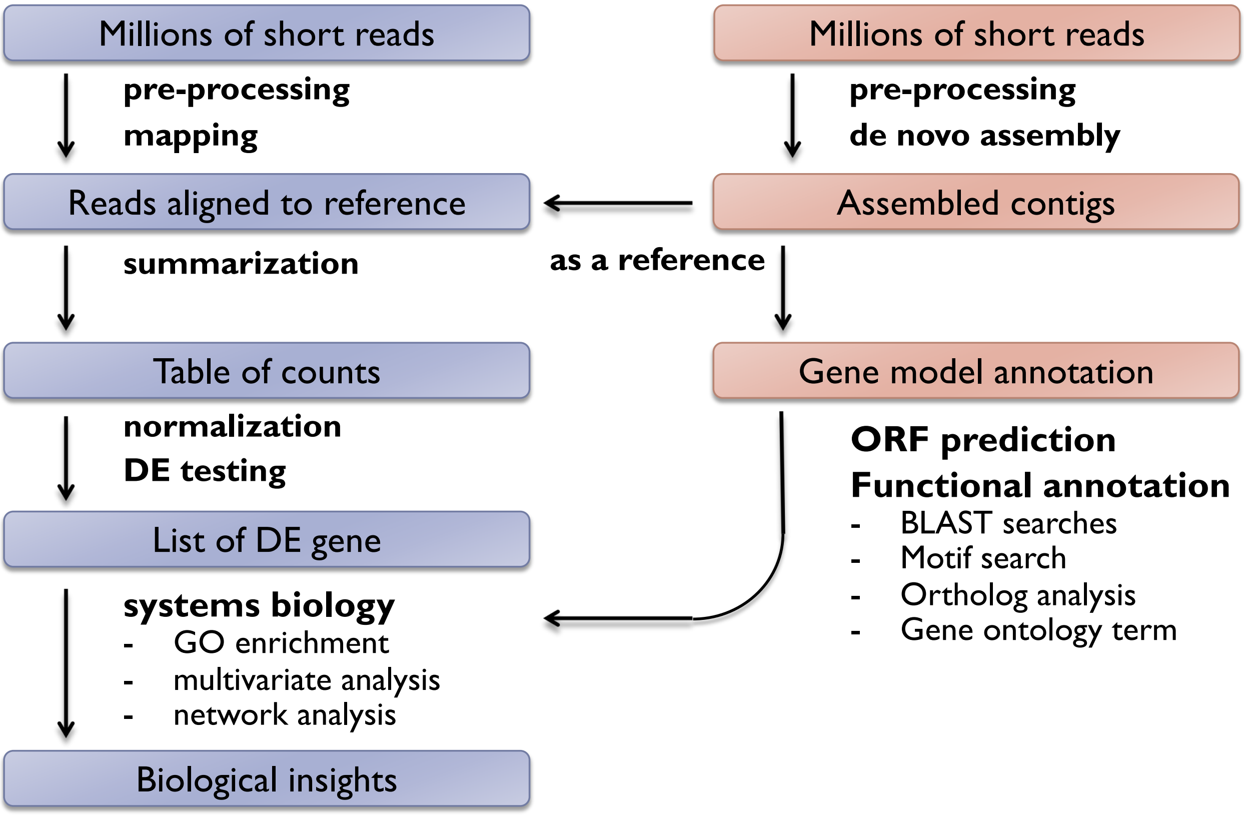
List of DE gene

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

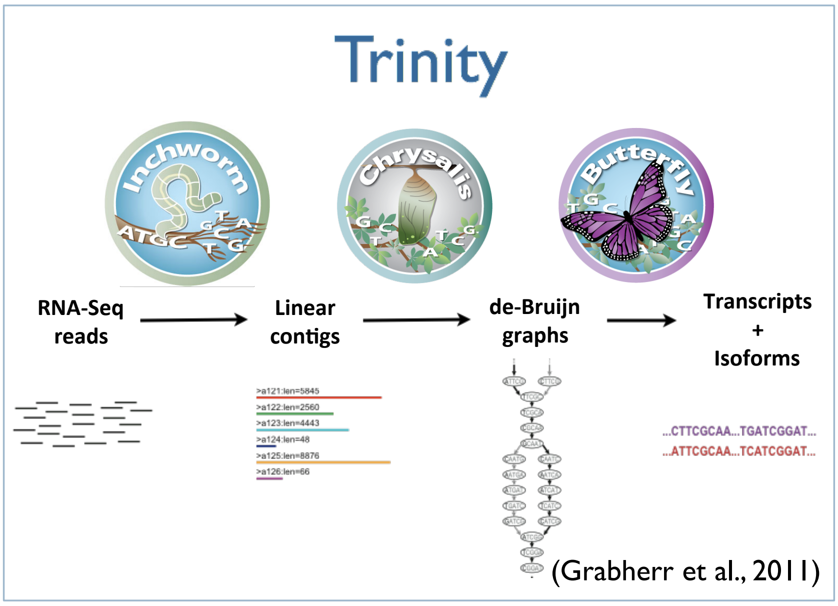
Functional annotation

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

de novo assemblers of RNA-seq

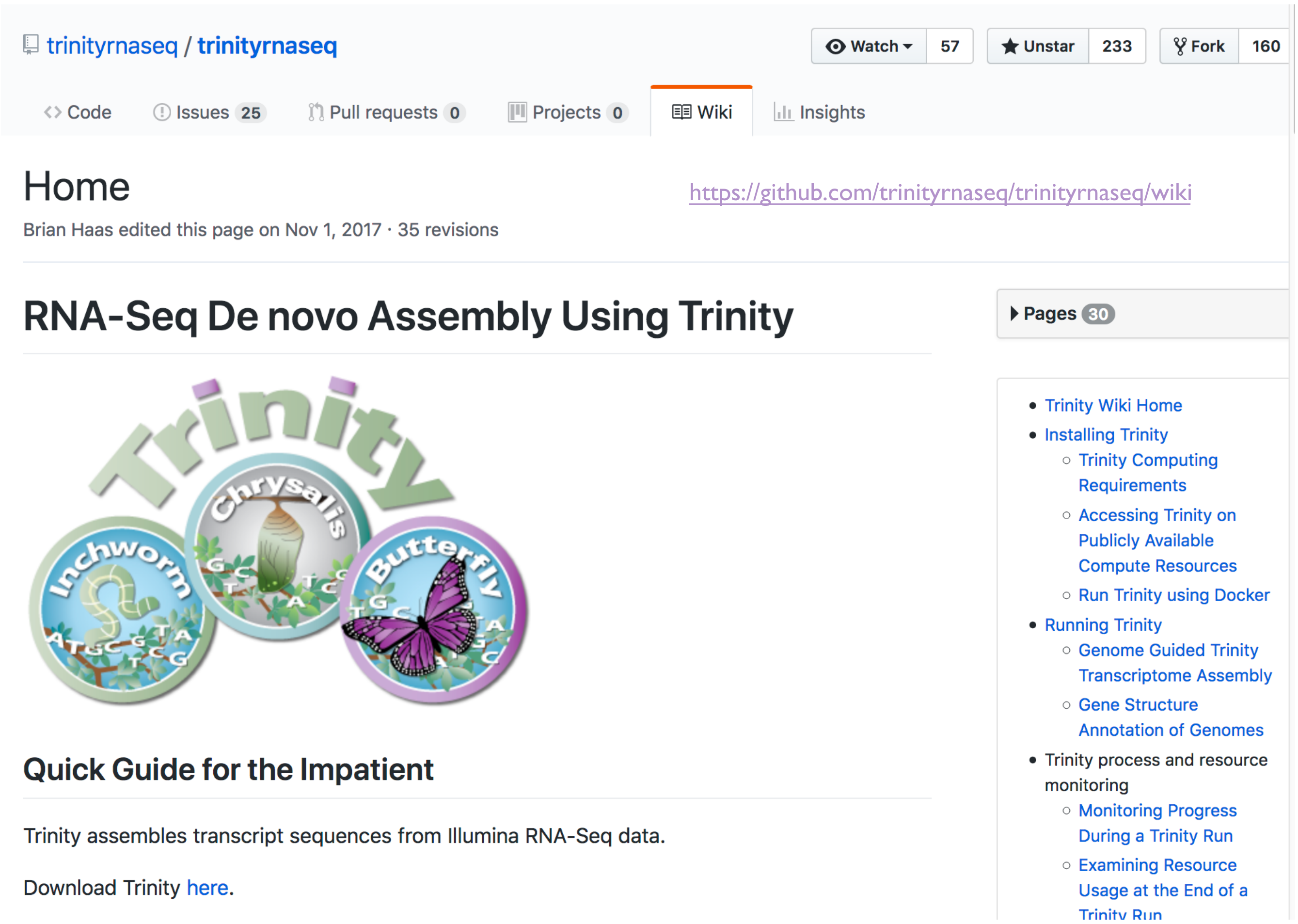
De novo assemblers use reads to assemble transcripts directly, which does not depend on a reference gnome.

- Trinity
- Oases
- TransAbyss
- EBARDenovo
- …



(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)

演習問題 ex9

Advanced

Clean up reference sequences

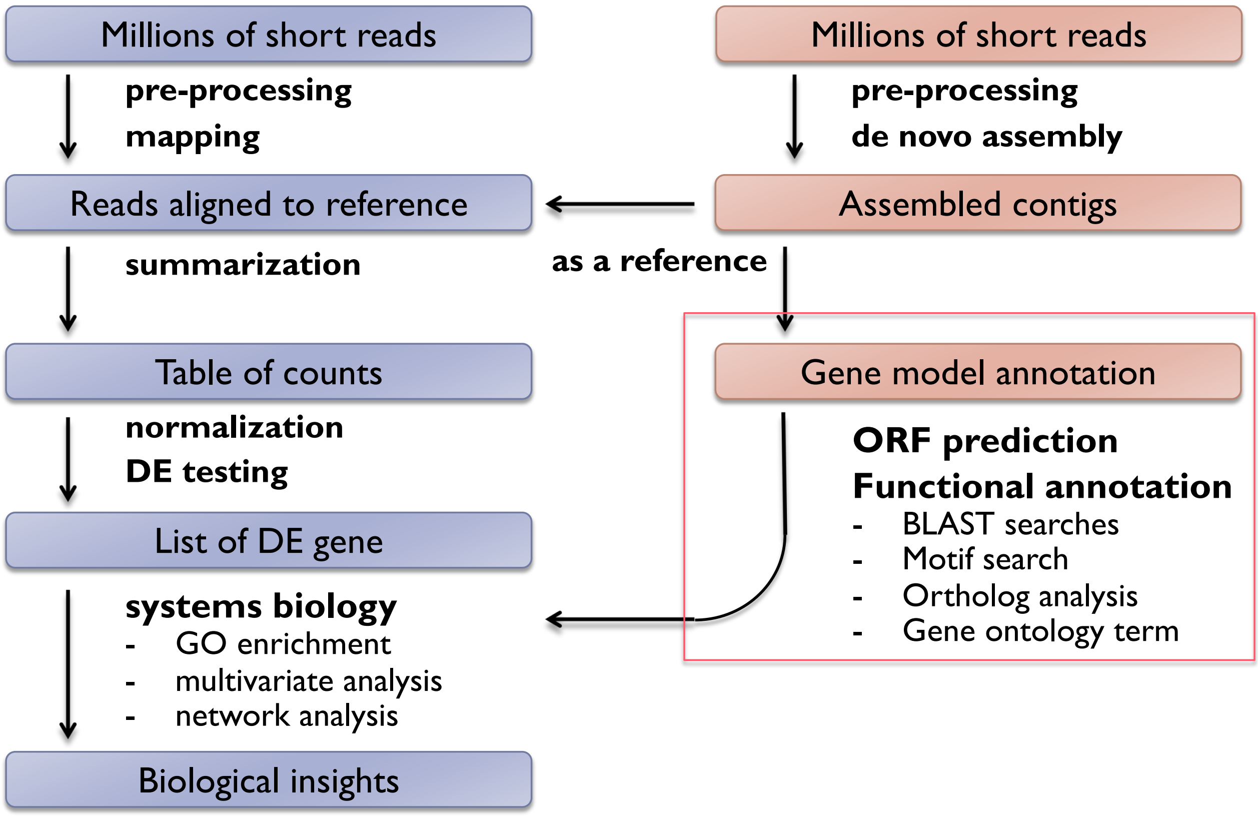
- An issue: Inflation of the number of Trinity contigs is often observed.
  - Trinity outputs splicing variants separately
  - Contaminations
  - Artifacts (bad contigs)
  - Incomplete contigs with very low expression.
- Solution
  - Filter out unwanted contigs.
  - Filter out very lowly expressed transcripts.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

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
```

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.