de novoシリーズ：第1回

非モデル生物のRNA-seq解析

～実験デザインから解析パイプラインまで～

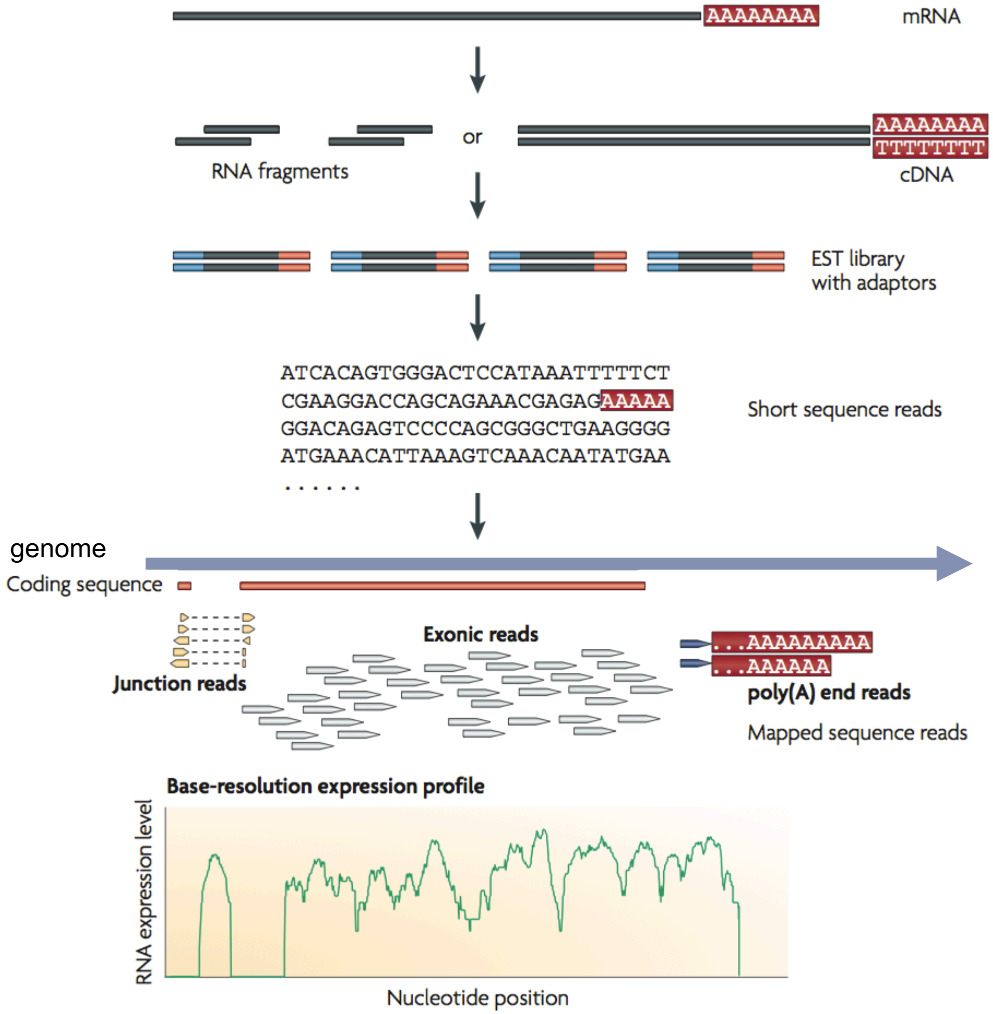
Shuji Shigenobu
重信　秀治

基礎生物学研究所
生物機能解析センター

April 3, 2013
Illumina Webinar Series

RNA-seq

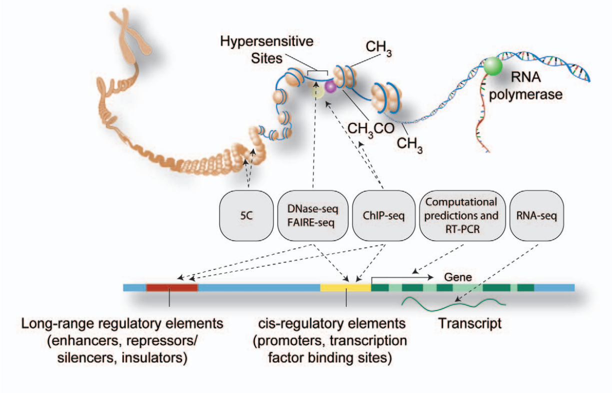
RNA-seq is a revolutionary tool for *transcriptomics* using deep-sequencing technologies.

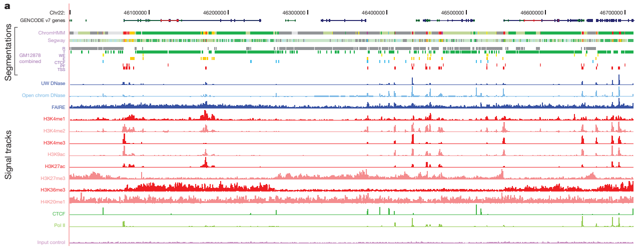


(Wang 2009 with modifications)

RNA-seq is unraveling complexities of eukaryotic transcriptomes in **model organisms**

- Differential expression
- Novel gene discovery
  - Coding and non-coding genes
- anti-sense transcripts
- RNA editing
- Novel splicing variants & fusion genes
- Allele-specific expression


Hypersensitive Sites
CH_3
RNA polymerase
CH_3CO
CH_3
5C
DNase-seq FAIRE-seq
ChIP-seq
Computational predictions and RT-PCR
RNA-seq
Gene
Long-range regulatory elements (enhancers, repressors/ silencers, insulators)
cis-regulatory elements (promoters, transcription factor binding sites)
Transcript

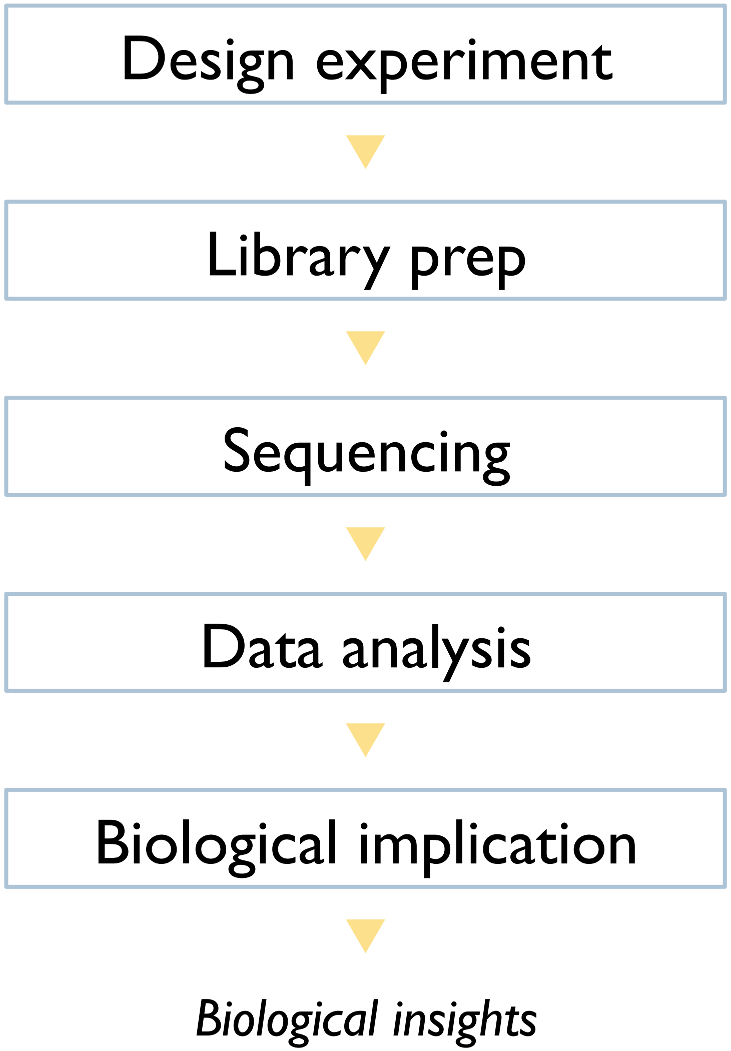


Is RNA-seq useful for **non-model species** without reference genome?

Yes!

- RNA-seq is very useful for organisms lacking sequenced genome.
- With recent technological advances, de novo strategy of RNA-seq works well.
- RNA-seq is much easier and cheaper than whole genome sequencing.

Workflow: NGS study



Design experiment

Library prep

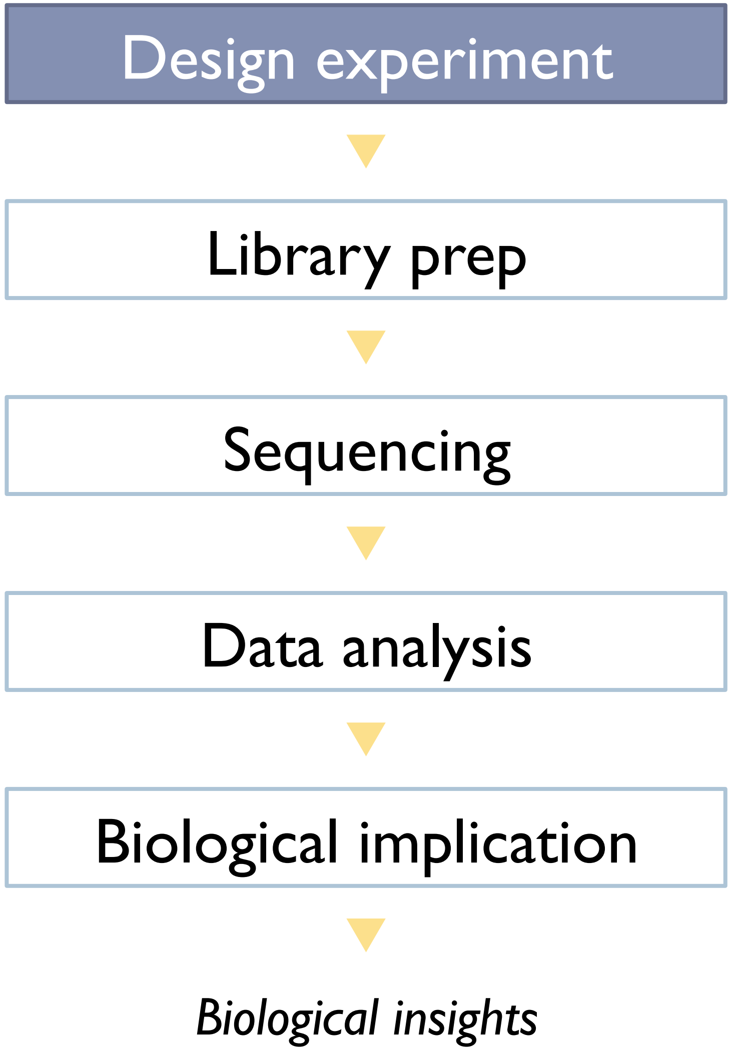
Sequencing

Data analysis

Biological implication

Biological insights

Workflow: NGS study



Design experiment

Library prep

Sequencing

Data analysis

Biological implication

Biological insights

Experimental design

- Issues to be considered in designing RNA-seq experiments.
  - You should define the **goal**.
  - Which **platform** do you choose?
  - **Depth**: How many reads do you need per sample?
  - **Length**: How long do you sequence?
  - **Paired-end** or single-end?
  - Method for **library construction**
    - Strand-specific?
    - Normalize?
  - How many biological **replicates**?
  - Pool RNA from multiple individuals or use a single individual?
  - Batch effect and lane effect.
  - **Informatics** strategy.

Experimental design

- Issues to be considered in designing RNA-seq experiments.
  - You should define the **goal**.
  - Which **platform** do you choose?
  - **Depth**: How many reads do you need per sample?
  - **Length**: How long do you sequence?
  - **Paired-end** or single-end?
  - Method for **library construction**
    - Strand-specific?
    - Normalize?
  - How many biological **replicates**?
  - Pool RNA from multiple individuals or use a single individual?
  - Batch effect and lane effect.
  - **Informatics** strategy.

Two major goals of RNA-seq

- Build gene catalogue
- Expression level quantification

Experimental design

- Issues to be considered in designing RNA-seq experiments.
  - You should define the **goal**.
  - Which **platform** do you choose?
  - **Depth**: How many reads do you need per sample?
  - **Length**: How long do you sequence?
  - **Paired-end** or single-end?
  - Method for **library construction**
    - Strand-specific?
    - Normalize?
  - How many biological **replicates**?
  - Pool RNA from multiple individuals or use a single individual?
  - Batch effect and lane effect.
  - **Informatics** strategy.

Choosing a platform

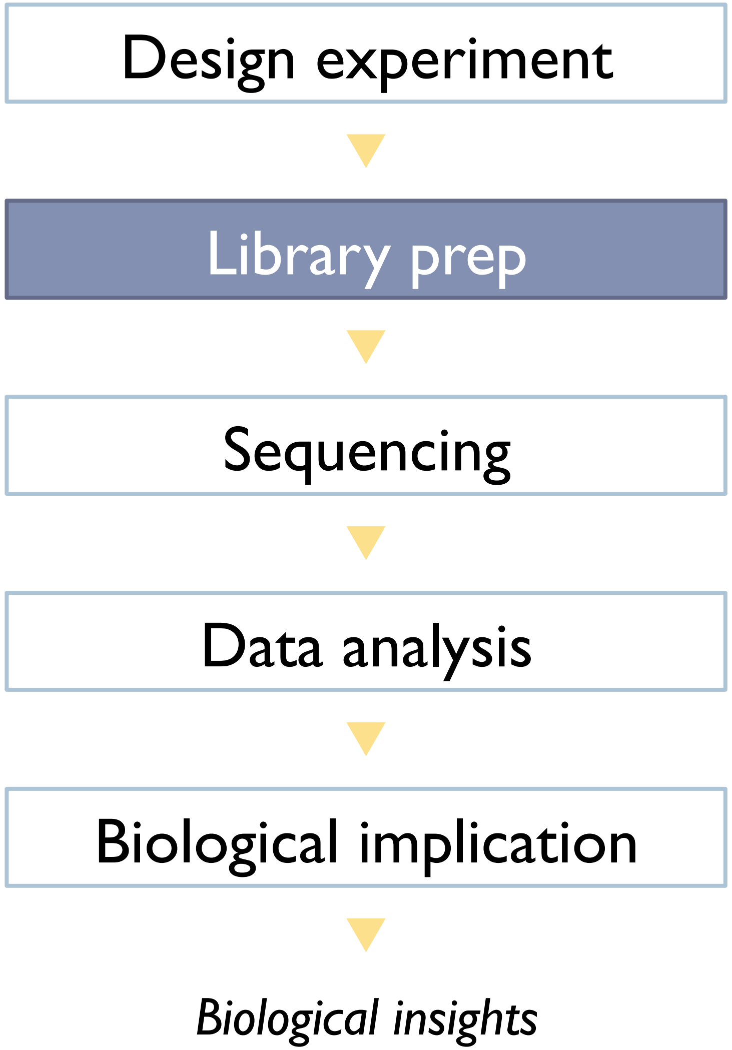
Illumina? 454? IonTorrent? PacBio? Or combined strategy?

- Use of **Illumina alone** is my recommendation as of today.

Experimental design

- Issues to be considered in designing RNA-seq experiments.
  - You should define the **goal**.
  - Which **platform** do you choose?
  - **Depth**: How many reads do you need per sample?
  - **Length**: How long do you sequence?
  - **Paired-end** or single-end?
  - Method for **library construction**
    - Strand-specific?
    - Normalize?
  - How many biological **replicates**?
  - Pool RNA from multiple individuals or use a single individual?
  - Batch effect and lane effect.
  - **Informatics** strategy.

Workflow: NGS study



Design experiment

Library prep

Sequencing

Data analysis

Biological implication

Biological insights

Library Prep: RNA extraction

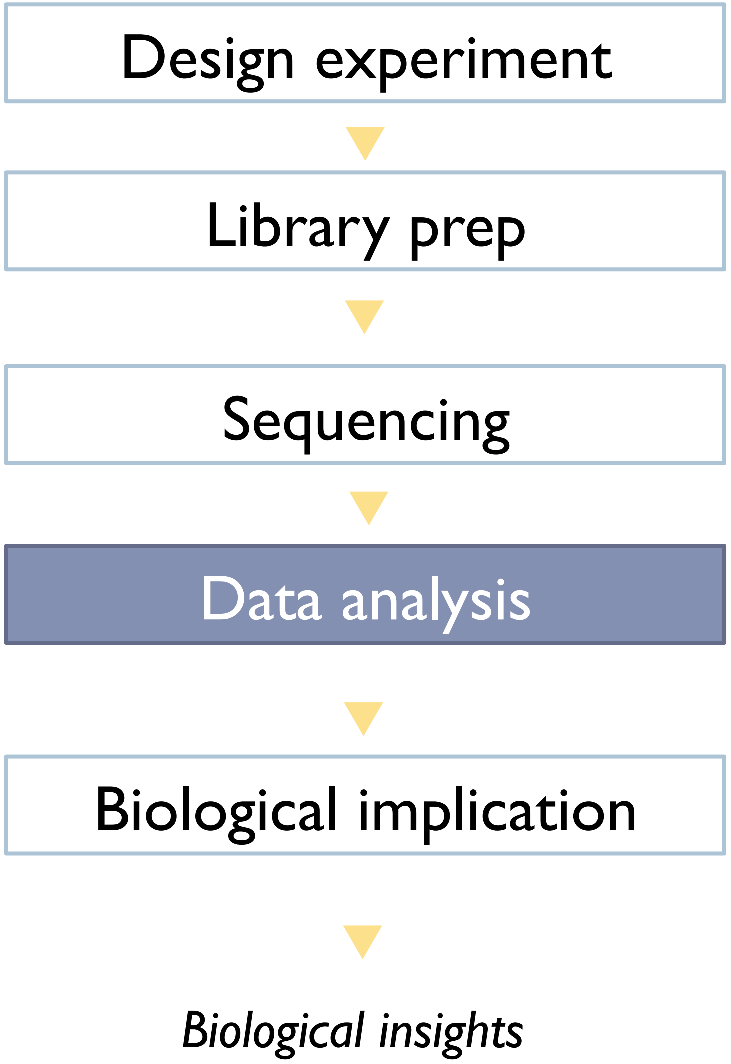
- RNA quality is the key to successful RNA-seq experiment
- RNA purification method: depends on the species and tissues.
- Poly A selection or rRNA depletion.
  - You may need pilot experiment for rRNA depletion kit, such as RiboMinus, because it was originally developed for model organisms.

Library construction method

- Illumina TruSeq RNA-seq prep kit
  - Normal kit
  - Strand-specific kit

- Third party kits for special uses
  - For small amount of RNA
  - Detect transcription start site

Workflow: NGS study



Design experiment

Library prep

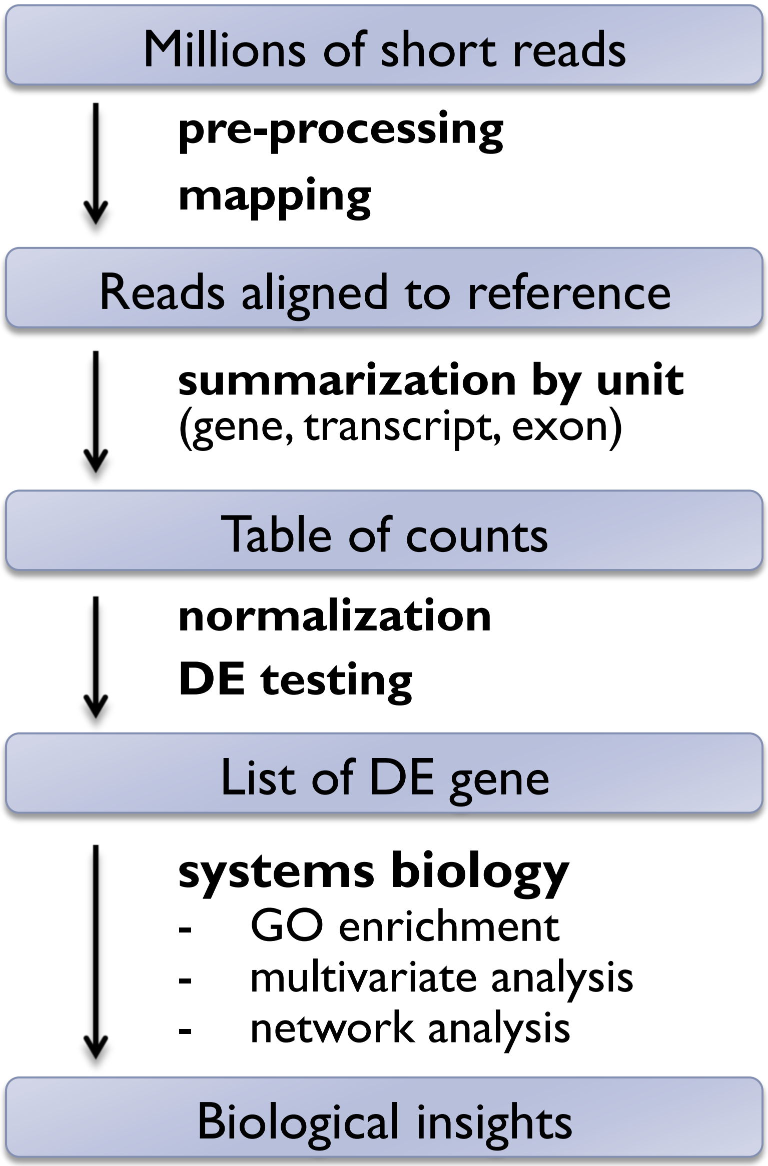
Sequencing

Data analysis

Biological implication

Biological insights

RNA-seq informatics workflow in model organisms



Millions of short reads

pre-processing

mapping

Reads aligned to reference

summarization by unit
(gene, transcript, exon)

Table of counts

normalization

DE testing

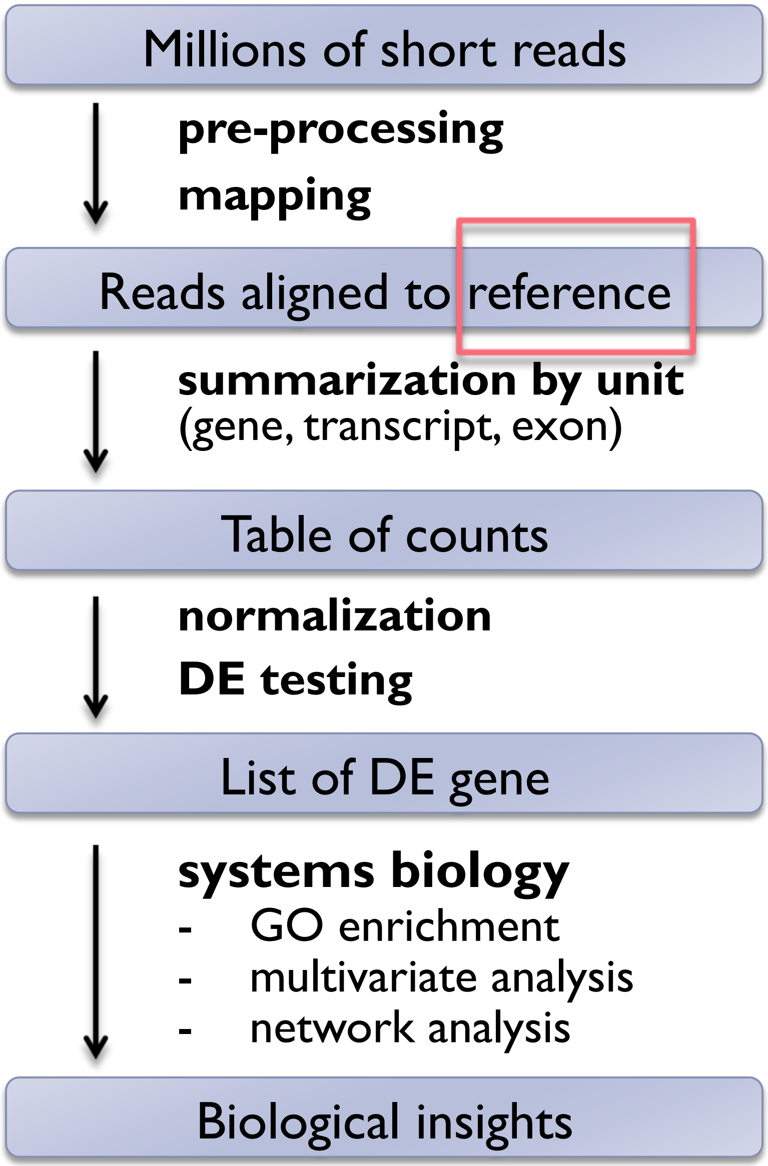
List of DE gene

systems biology

- GO enrichment
- multivariate analysis
- network analysis

Biological insights

RNA-seq informatics workflow in model organisms



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

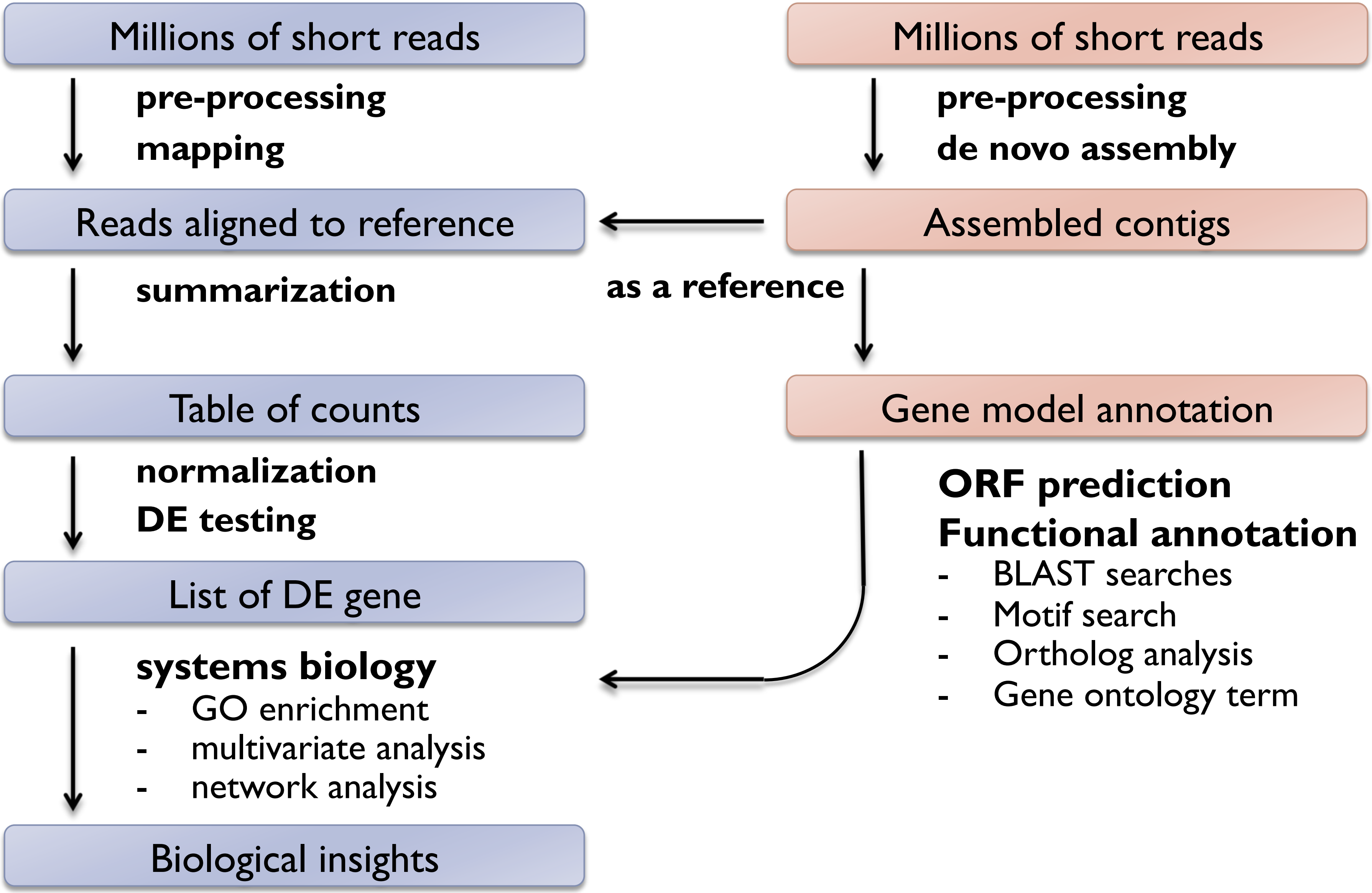
systems biology

- 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

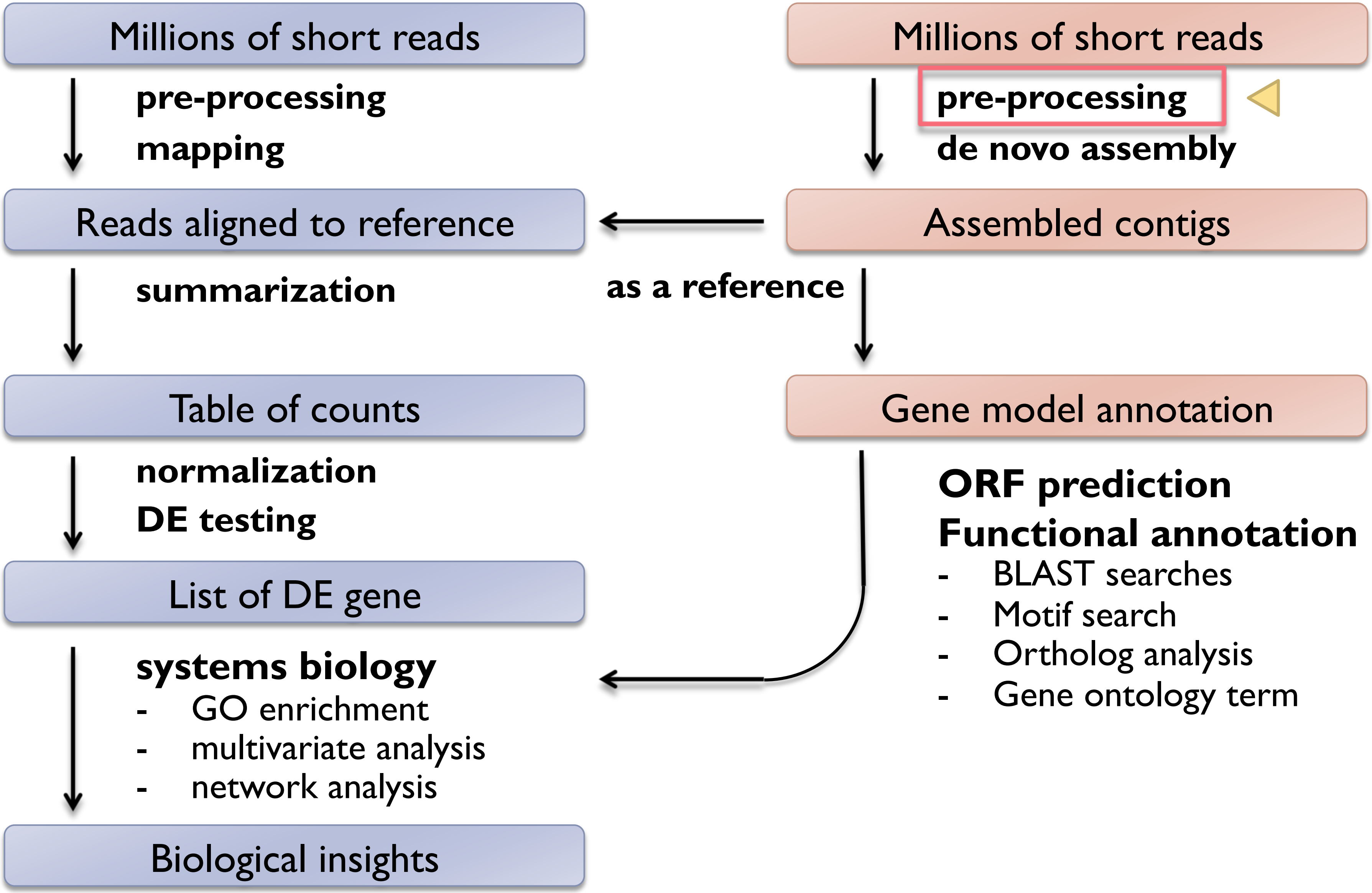
Gene model annotation

ORF prediction

Functional annotation

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

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

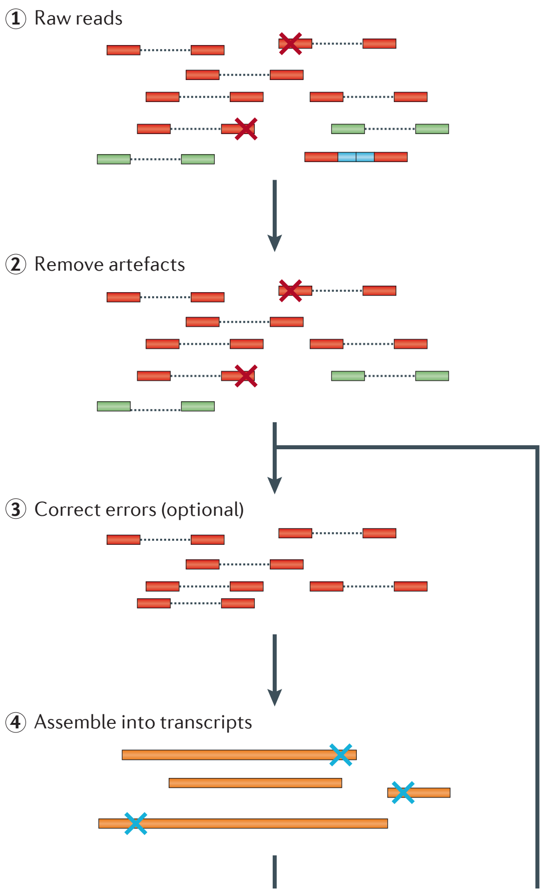
Functional annotation

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

Pre-processing of short reads

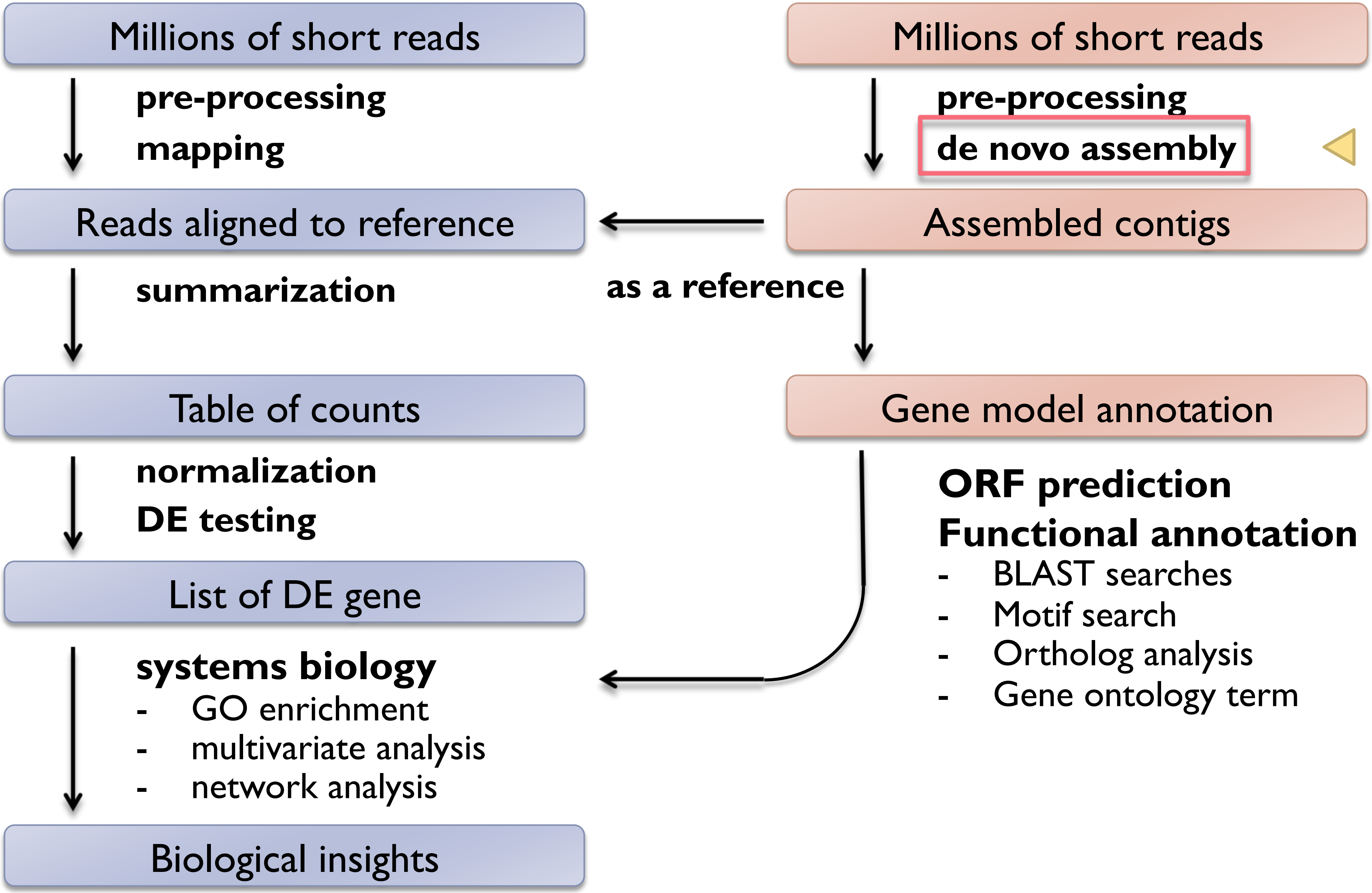
- Filter or trim by base quality
- Remove artifacts
  - adaptors
  - low complexity reads
  - PCR duplications (optional)
- Remove rRNA and other contaminations (optional)
- Sequence error correction (optional)

Suggestion: Pre-processing is strongly recommended for de novo assembly.



Martin et al (2011) *Nat Rev Genet*

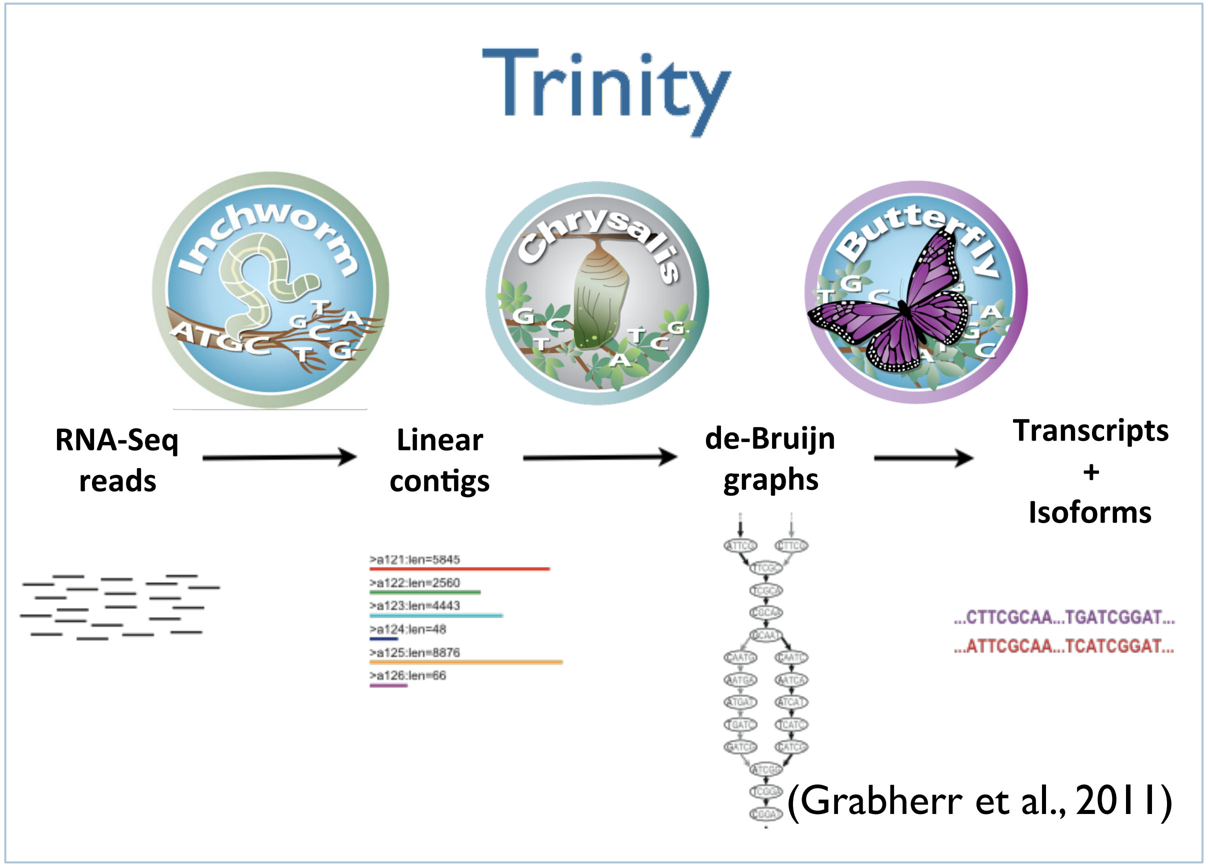
RNA-seq analysis pipeline (*de novo* strategy)



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
- …



(Grabherr et al., 2011)

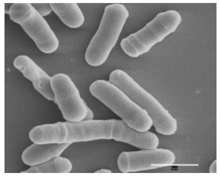
http://trinityrnaseq.sourceforge.net/

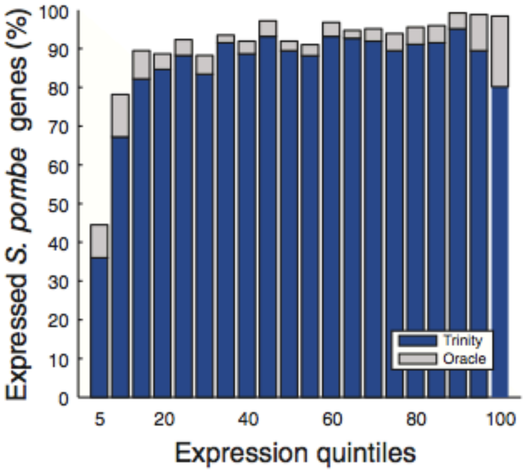
Transcript reconstruction by expression quaintile using Trinity

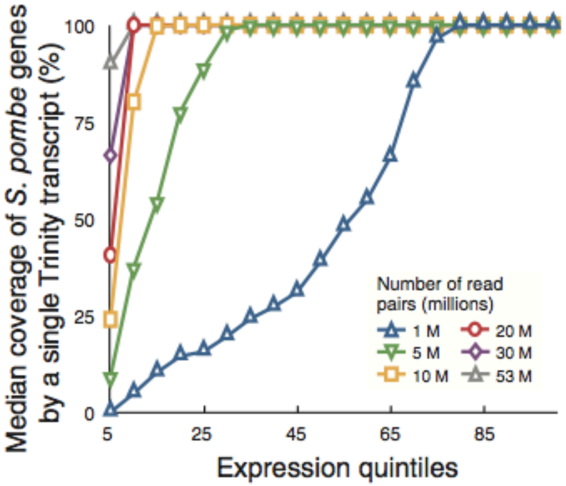
Full-length Reconstruction

Reconstruction vs. Sequencing Depth

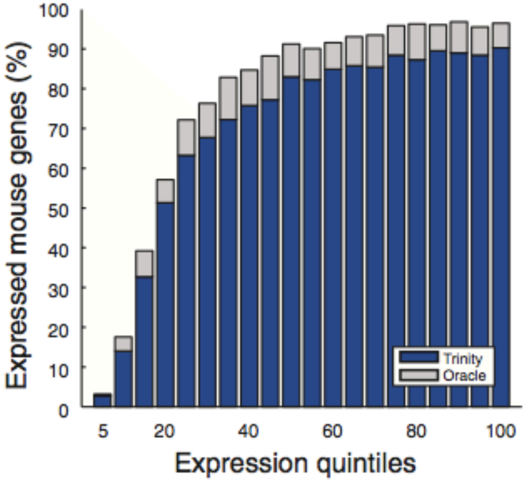
S. pombe

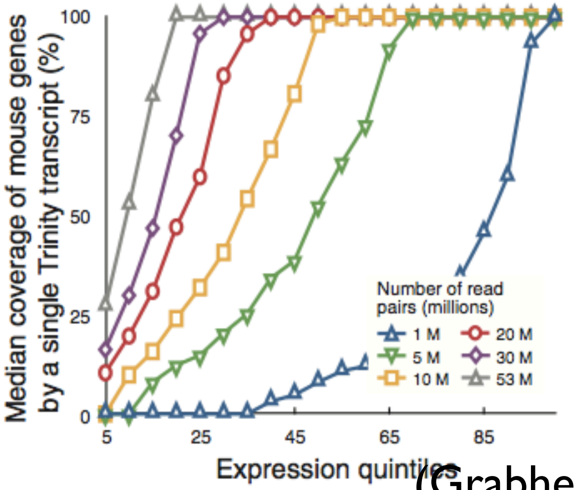






Mouse





(Grabherr et al. 2011)

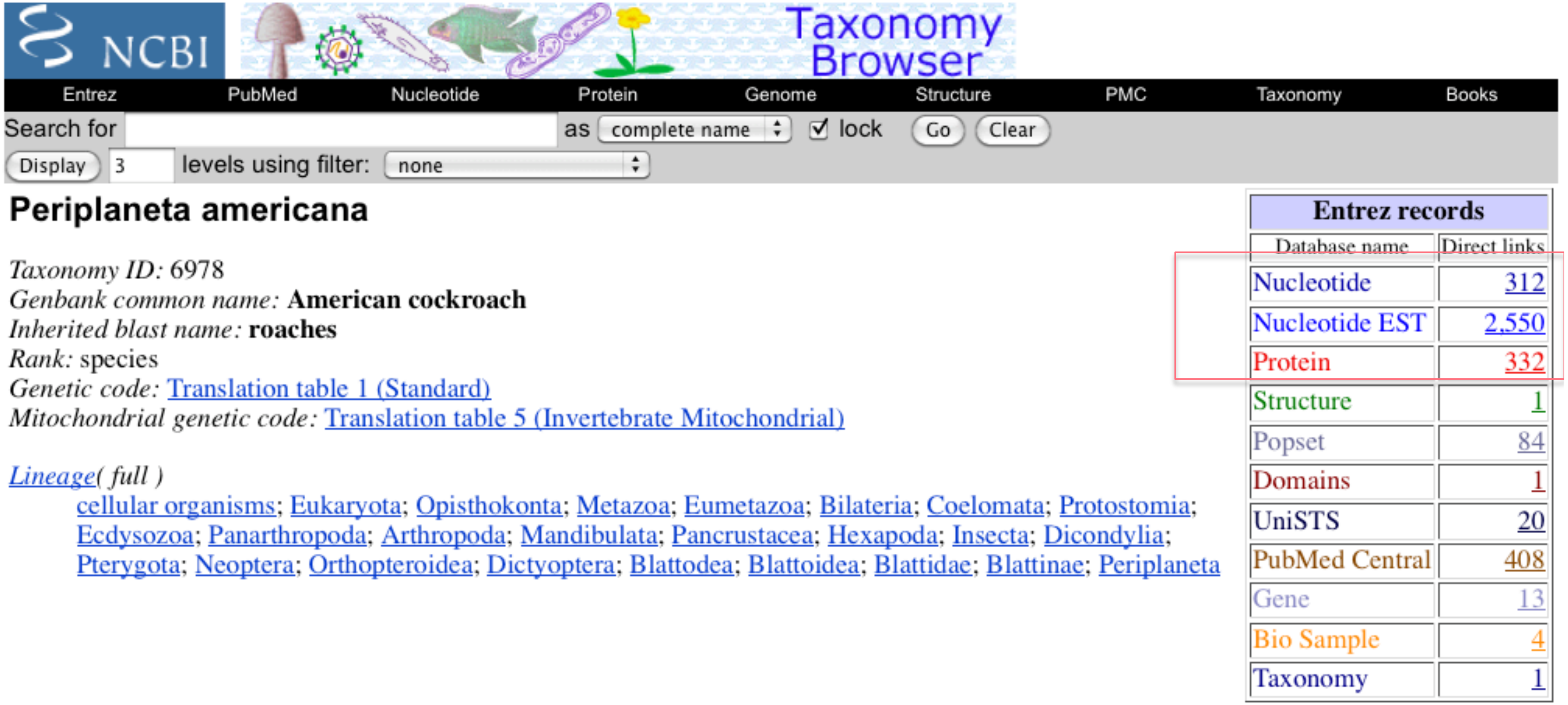
our example

Cockroach RNA-seq

- Motivation:
  - Hygienic pest
  - Developmental biology
    - appendage regeneration
  - Social biology
    - comparison with termites
  - Neuroscience
  - Symbiosis with bacteria

Periplaneta americana
ワモンゴキブリ Photo:wikipedia

(Collaboration with Miura Lab of 北大)



Periplaneta americana

Taxonomy ID: 6978
Genbank common name: **American cockroach**
Inherited blast name: **roaches**
Rank: species
Genetic code: Translation table 1 (Standard)
Mitochondrial genetic code: Translation table 5 (Invertebrate Mitochondrial)

Lineage(*full*)
cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Coelomata; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Orthopteroidea; Dictyoptera; Blattodea; Blattoidea; Blattidae; Blattinae; Periplaneta

Entrez records

Database name	Direct links
Nucleotide	312
Nucleotide EST	2,550
Protein	332
Structure	1
Popset	84
Domains	1
UniSTS	20
PubMed Central	408
Gene	13
Bio Sample	4
Taxonomy	1

Little genetic / genomic information is available for cockroaches
One of the reason is the large genome size

Cockroach RNA-seq

Periplaneta americana

- 6 libraries [Illumina TruSeq]
- Multiplexed Sequencing [HiSeq2000]
  - Paired-end 101+101bp (HiSeq ver.2 half lane)

Embryos	Young larvae	Late larva ♀	Late larva ♂	Adult ♀	Adult ♂
9.6M	9.4M	9.1M	10.0M	8.1M	9.8M

55.8M read pairs (11.2G bp)

De novo assembly with Trinity

146,172 contigs (≈ isoforms)
90,837 components (≈ genes)

(Shigenobu, Hayashi and Miura, in prep)

Assembly Evaluation

- Assembly statistics
  - (example: our cockroach RNA-seq)
    - # components: 90,473
    - Mean: 772.2 base
    - N50: 1384 base
    - Total bases: 69.9 Mb
- Quality control
  - No commonly accepted methods for de novo RNA-seq assembly.
  - Proposed metrics:
    - accuracy, completeness, contiguity, chimerism and variant resolution (Martin and Wang, 2011)
- Find artifacts and contaminations

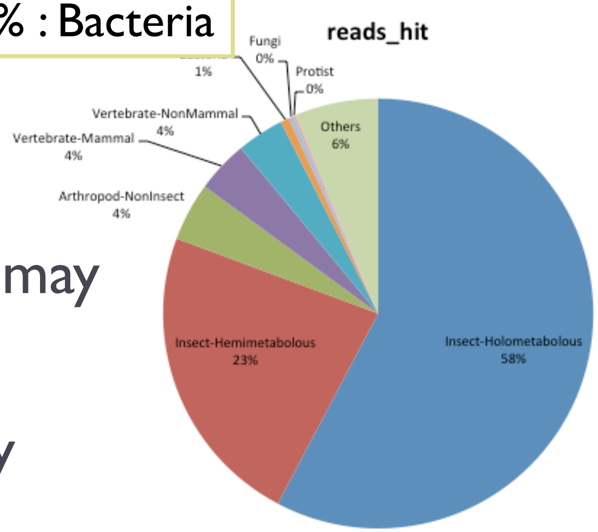
our example

Bonus from RNA-seq “Contamination”

- Full-length rRNA
  - Low level rRNA contamination reads (~0.5%) are enough to recapitulate complete rRNA
  - 7,242bp rRNA obtained (Complete18S+28S) [New!]

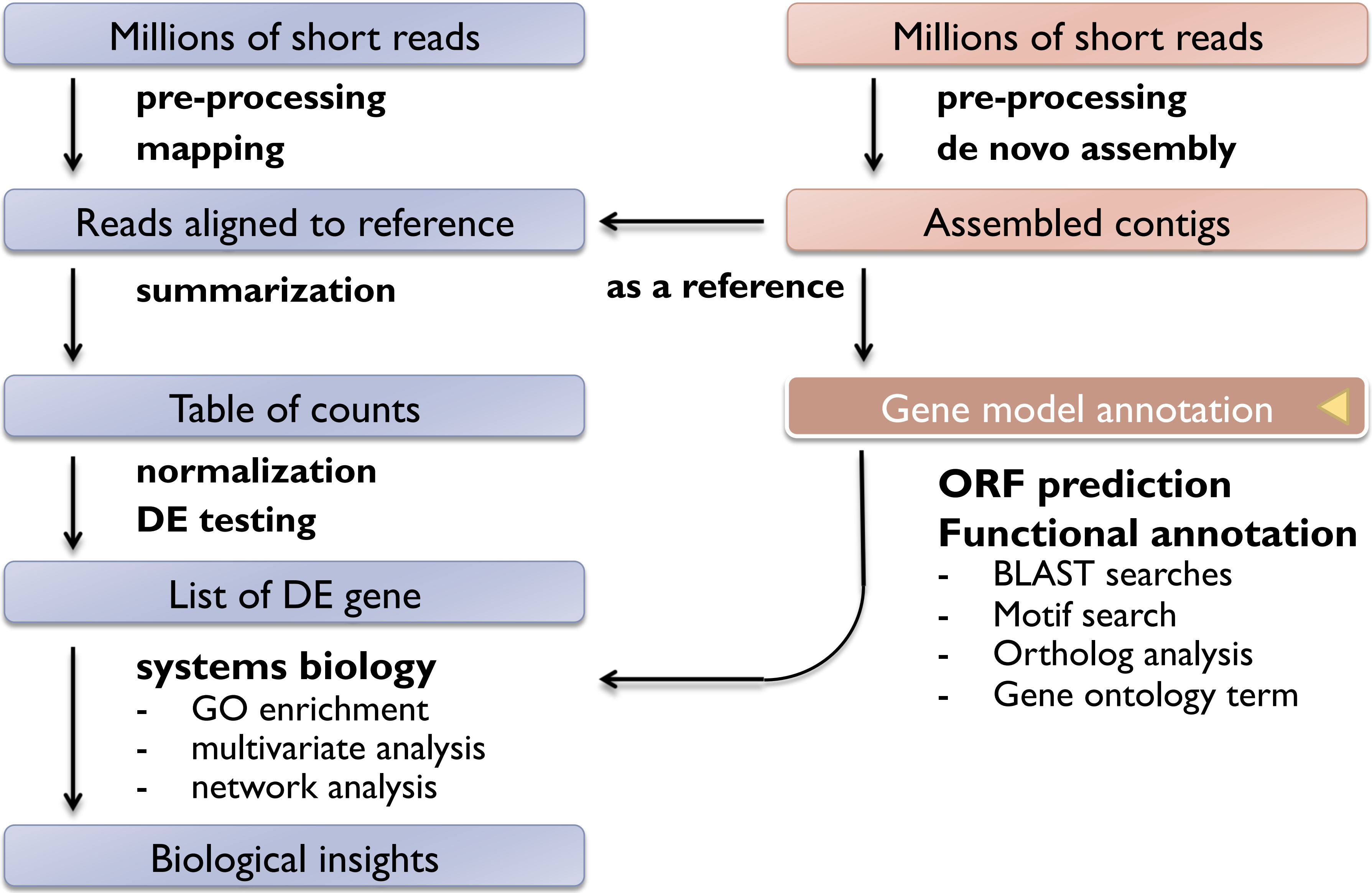
- Symbiont RNAs
  - AT-rich bacterial transcripts remain.
  - Some are just contamination, while some may be important partners, e.g. symbionts.
  - 80 Genes of ***Blattabacterium*** (obligatory endosymbiont of cockroach) found.

0.65% : Bacteria



BLAST nr tophit taxonomy

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.
    - Don’t worry. Frame shifts do not occur so often in Illumina.

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

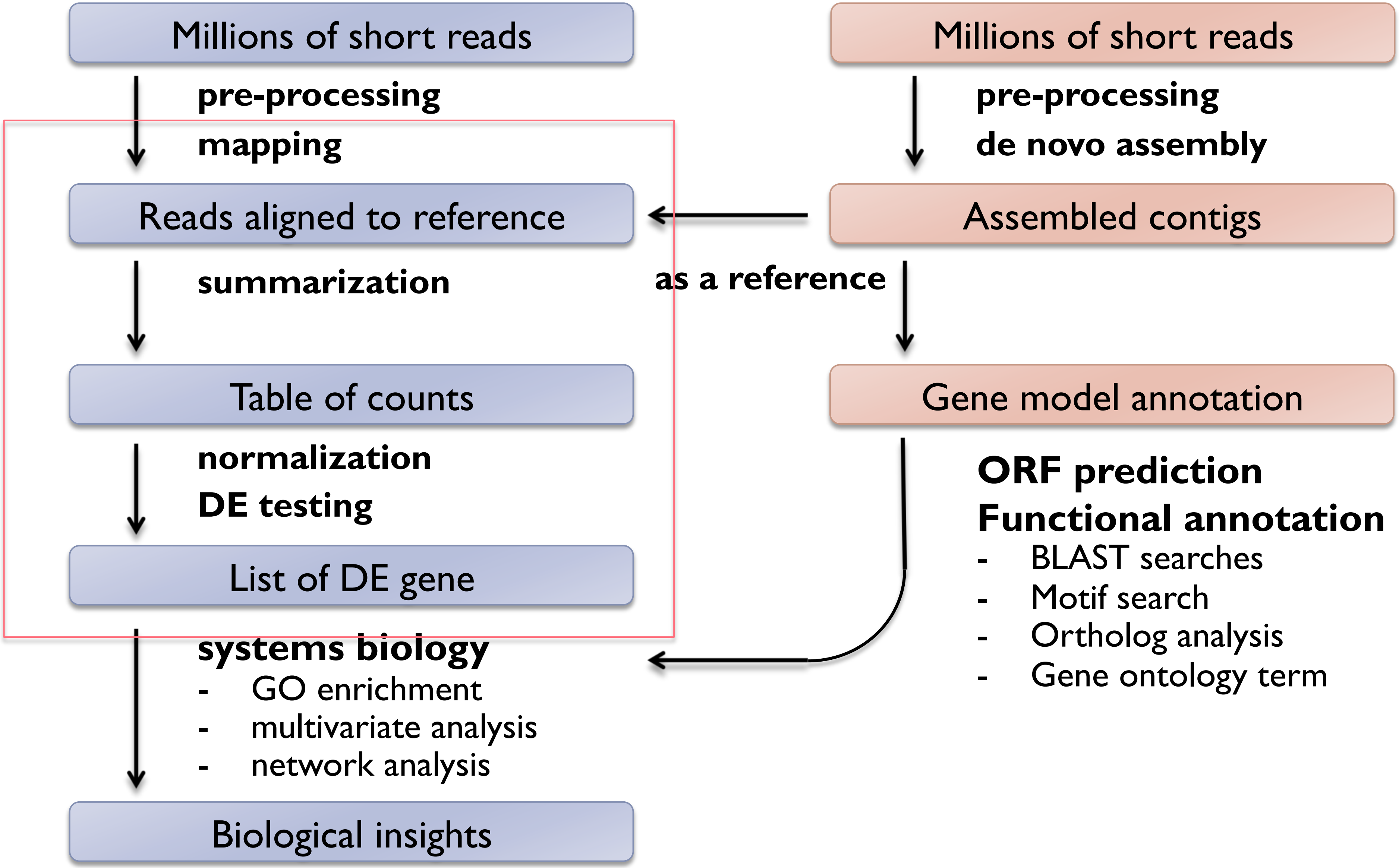
our example

Cockroach RNA-seq

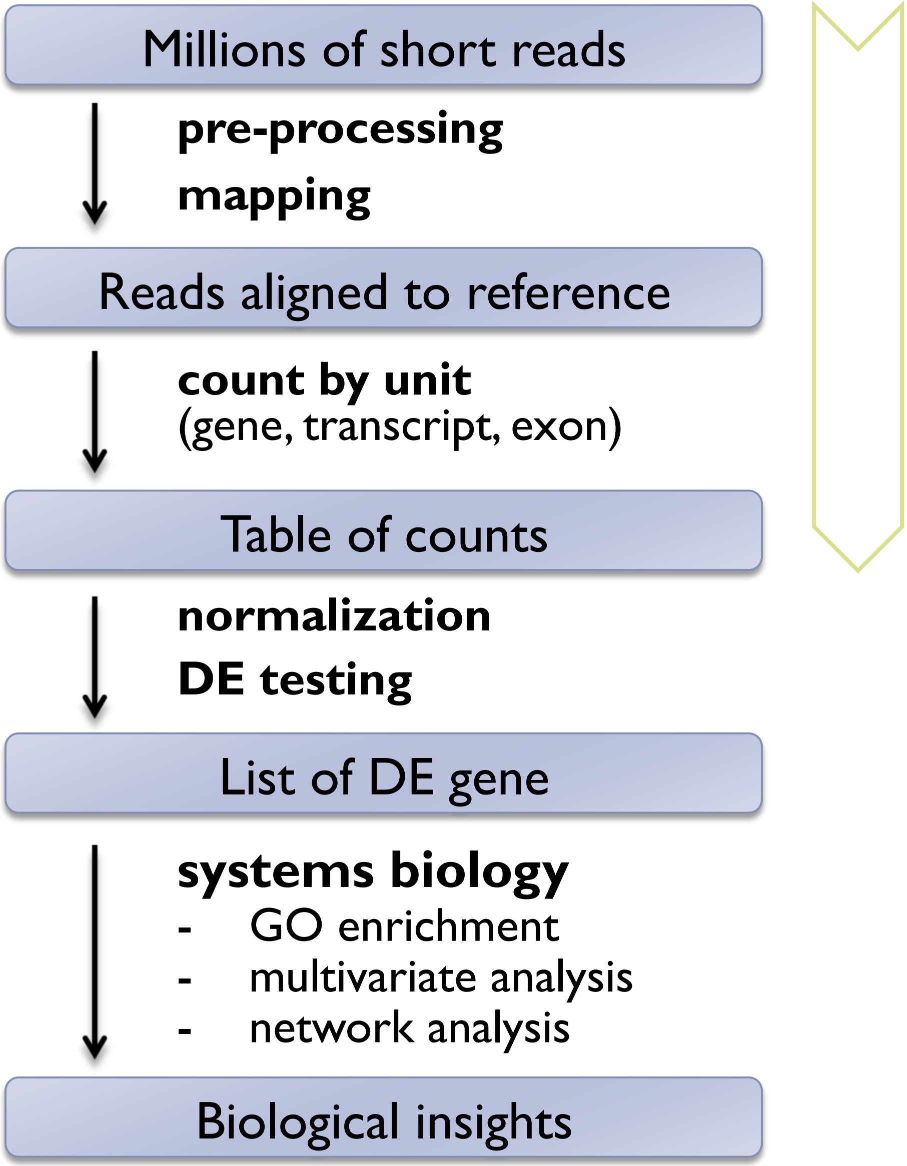
- ORF prediction
  - 28,649 (> 50aa)

- Gene repertoire in comparison with other insects
  - 16,826 show similarity w/ 7539 *D. melanogaster* genes [54.7% of Dmel gene set]
  - 18,233 show similarity w/ 7149 *Pediculus humanus* genes [66.3% of Phum gene set]
  - 25,524 (89.0%) represent 9,419 arthropod ortholog groups. (based on OrthoDB)

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



RNA-seq analysis pipeline for DE



Millions of short reads

pre-processing

mapping

Reads aligned to reference

count by unit
(gene, transcript, exon)

Table of counts

normalization

DE testing

List of DE gene

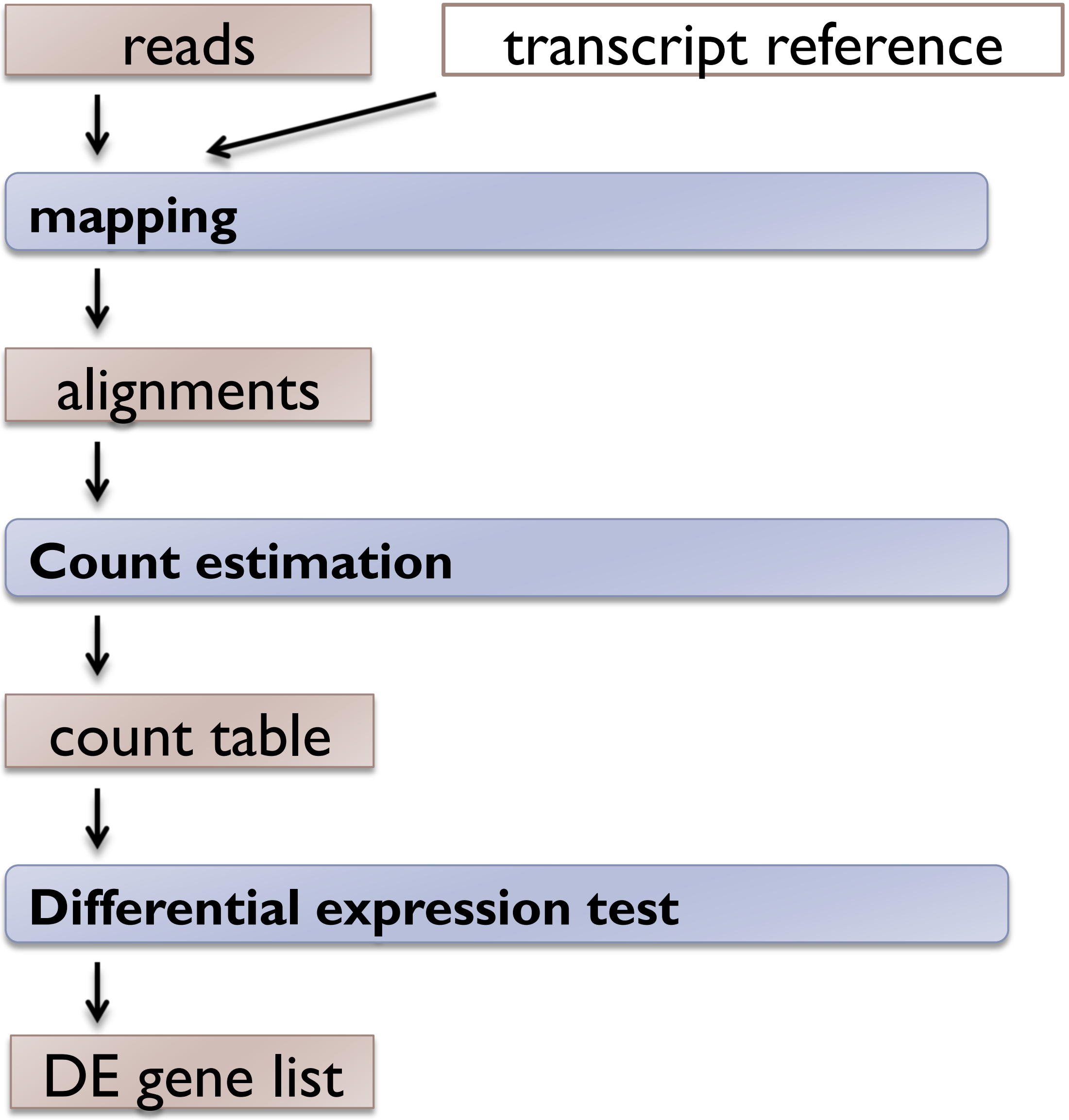
systems biology

- GO enrichment
- multivariate analysis
- network analysis

Biological insights

data type	**format**
sequences	fastq
alignment	SAM/BAM
table	text (tab delimited)

Differential expression analysis



reads

transcript reference

mapping

alignments

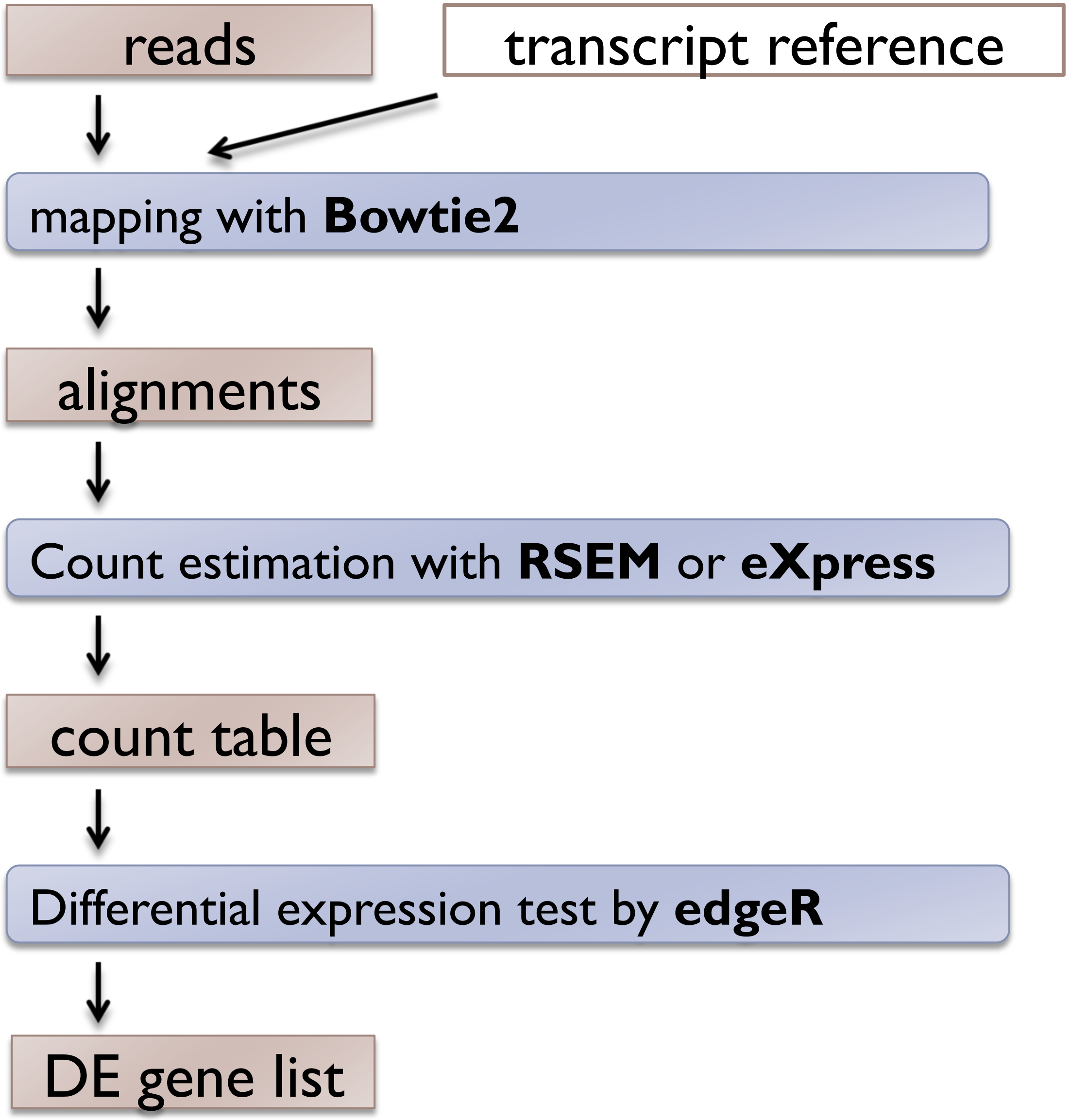
Count estimation

count table

Differential expression test

DE gene list

Differential expression analysis



reads

transcript reference

mapping with **Bowtie2**

alignments

Count estimation with **RSEM** or **eXpress**

count table

Differential expression test by **edgeR**

DE gene list

Mapping – alignment software

Many aligners have been developed for short read mapping

- Reference = Transcripts:

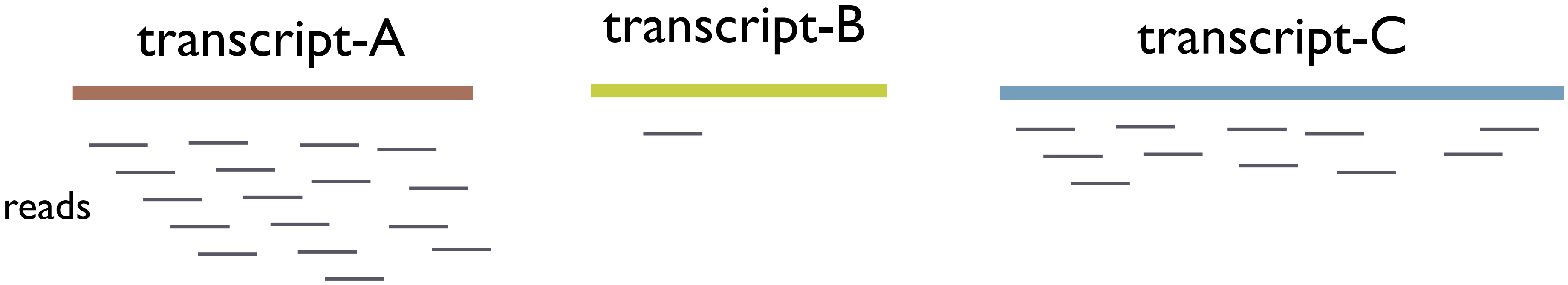
  short read mapper (unspliced read aligner) is used

  - **Bowtie2** – basic mapping to reference sequence

    http://bowtie-bio.sourceforge.net/bowtie2/index.shtml

    others – BWA, SOAP2, PerM, SHRiMP, BFAST, ELAND

Count Reads by Transcript



- The simplest way: just count reads by contig.

But…

- Multimapping issue should be considered.

Estimate Abundance

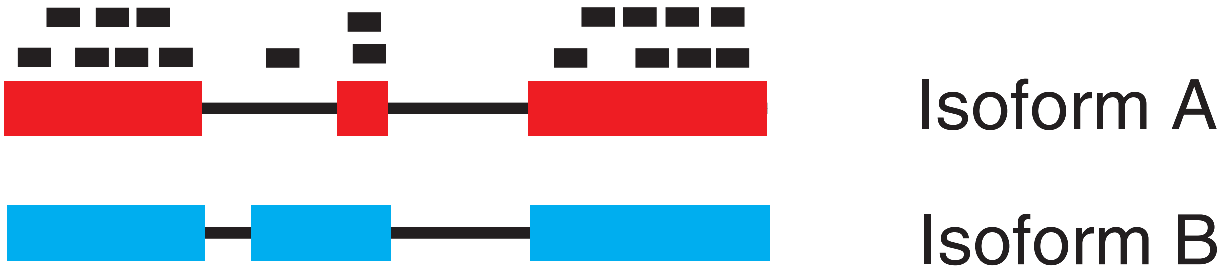
- **Multimapping issues**
  - Isoforms
  - Repetitive sequences

- Mapping ambiguity should be taken into consideration.

Estimate Abundance

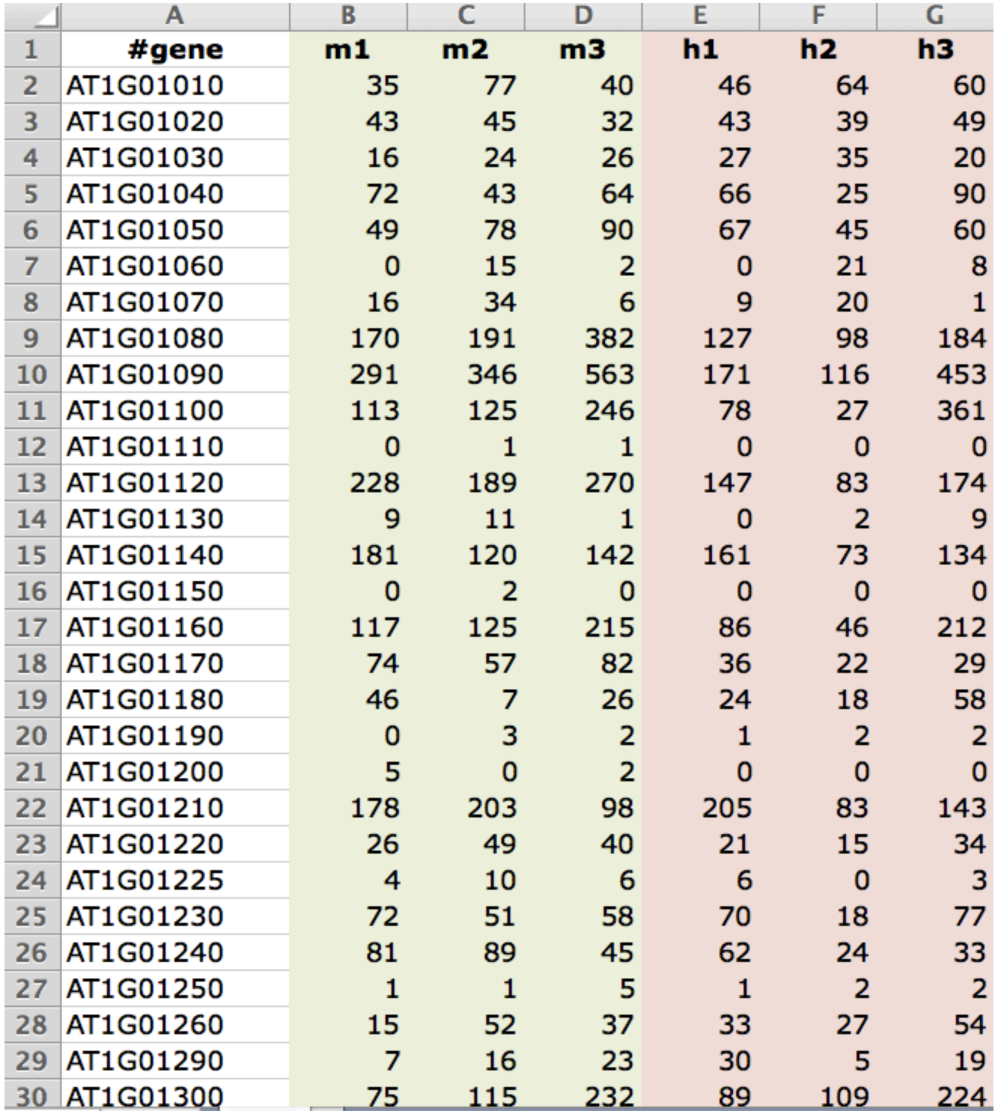
- **Multimapping issues**
  - Isoforms ← important in working with Trinity output
  - Repetitive sequences

- Mapping ambiguity should be taken into consideration.



- Software: RSEM and eXpress (EM algorithm) conditions

genes



	A	B	C	D	E	F	G
1	#gene	m1	m2	m3	h1	h2	h3
2	AT1G01010	35	77	40	46	64	60
3	AT1G01020	43	45	32	43	39	49
4	AT1G01030	16	24	26	27	35	20
5	AT1G01040	72	43	64	66	25	90
6	AT1G01050	49	78	90	67	45	60
7	AT1G01060	0	15	2	0	21	8
8	AT1G01070	16	34	6	9	20	1
9	AT1G01080	170	191	382	127	98	184
10	AT1G01090	291	346	563	171	116	453
11	AT1G01100	113	125	246	78	27	361
12	AT1G01110	0	1	1	0	0	0
13	AT1G01120	228	189	270	147	83	174
14	AT1G01130	9	11	1	0	2	9
15	AT1G01140	181	120	142	161	73	134
16	AT1G01150	0	2	0	0	0	0
17	AT1G01160	117	125	215	86	46	212
18	AT1G01170	74	57	82	36	22	29
19	AT1G01180	46	7	26	24	18	58
20	AT1G01190	0	3	2	1	2	2
21	AT1G01200	5	0	2	0	0	0
22	AT1G01210	178	203	98	205	83	143
23	AT1G01220	26	49	40	21	15	34
24	AT1G01225	4	10	6	6	0	3
25	AT1G01230	72	51	58	70	18	77
26	AT1G01240	81	89	45	62	24	33
27	AT1G01250	1	1	5	1	2	2
28	AT1G01260	15	52	37	33	27	54
29	AT1G01290	7	16	23	30	5	19
30	AT1G01300	75	115	232	89	109	224

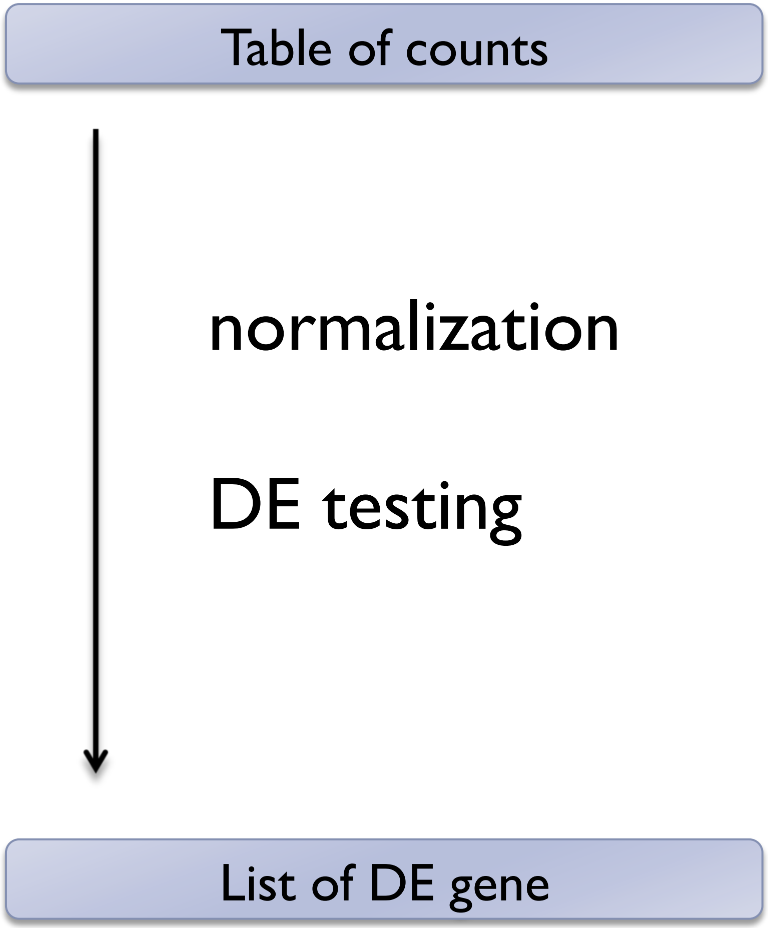


Table of counts

normalization

DE testing

List of DE gene

DE: Differential Expression

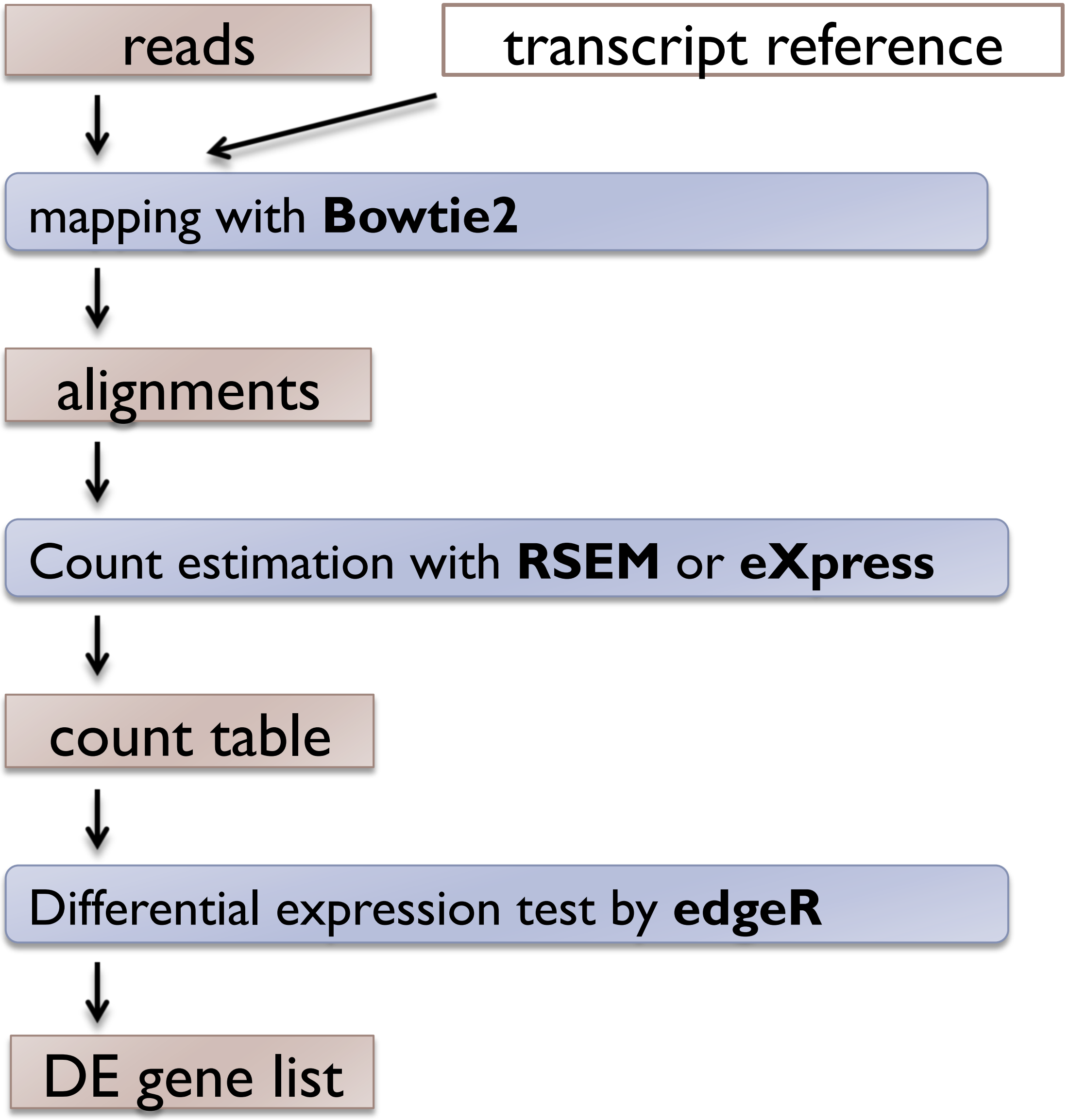
Software for RNA-seq DE analysis

- Many software available
  - **edgeR**
  - Genominator
  - DESeq
  - DEGSeq
  - baySeq
  - NBPSeq
  - TCC
  - …

edgeR

- A Bioconductor package for differential expression analysis of digital gene expression data
- **Model**: An over dispersed Poisson model, negative binomial (NB) model is used
- **Normalization**: TMM method (trimmed mean of M values) to deal with composition effects
- **DE test**: exact test and generalized linear models (GLM)

Differential expression analysis



reads

transcript reference

mapping with **Bowtie2**

alignments

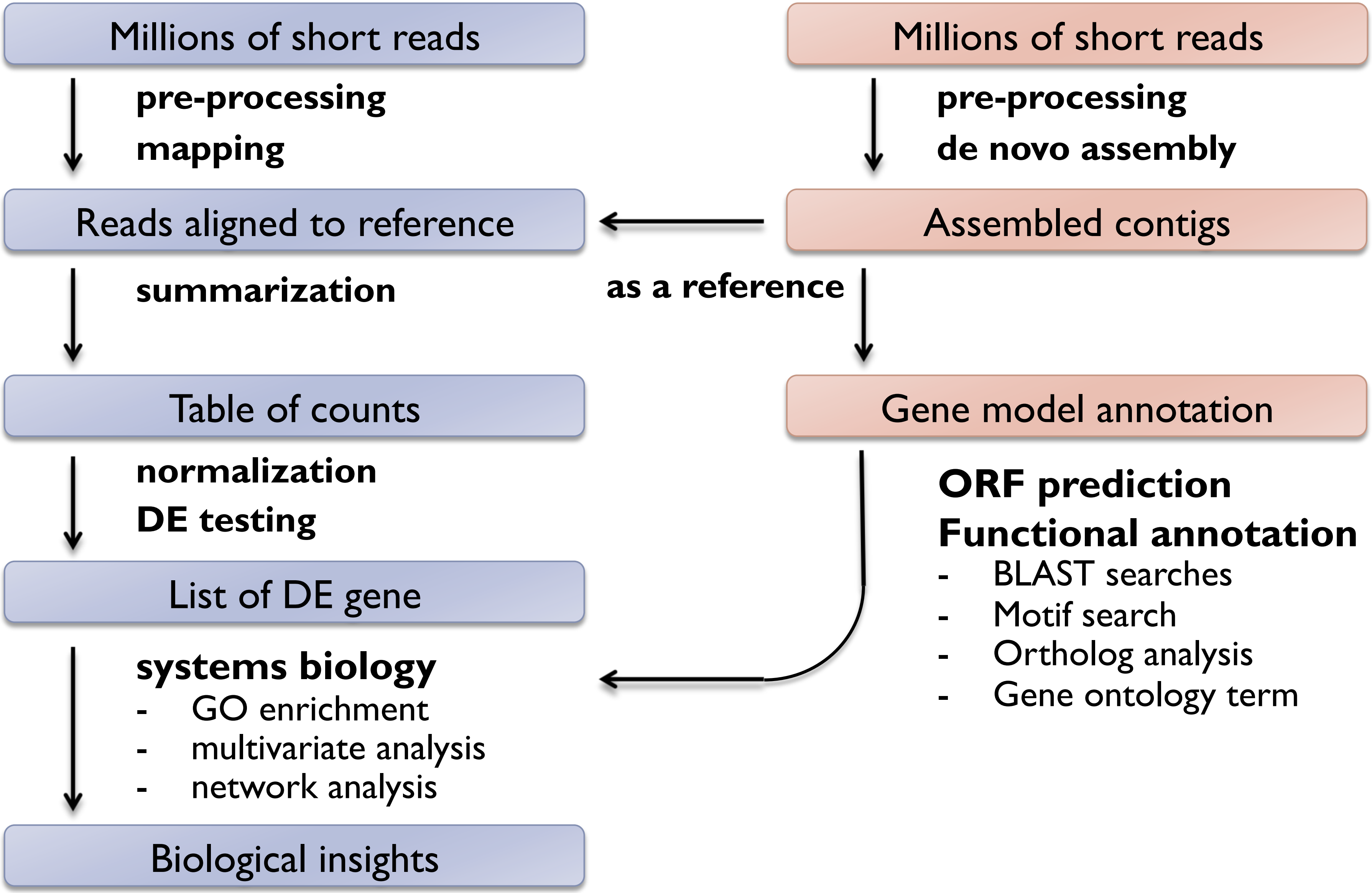
Count estimation with **RSEM** or **eXpress**

count table

Differential expression test by **edgeR**

DE gene list

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



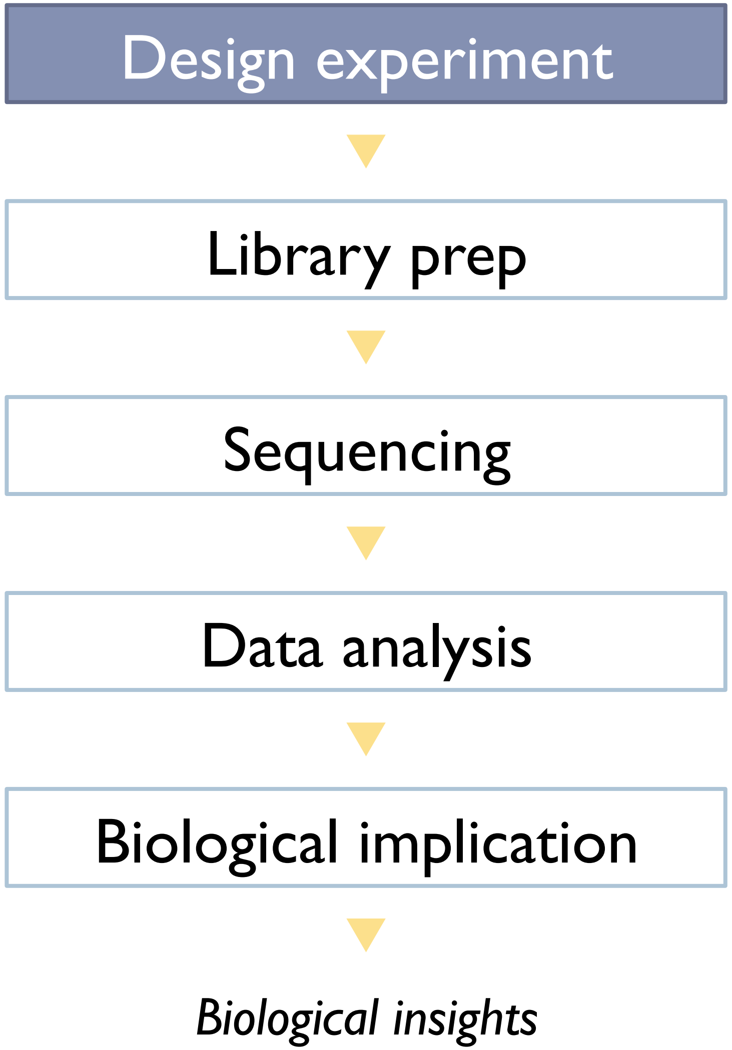
Beyond transcriptome: Other applications of *de novo* RNAseq assembly

- **Proteomics**:
  Build proteome database for peptide mass fingerprinting

- **Genomics**:
  SNP identification

- **“Homolog” cloning**:
  Alternative to “degenerate PCR” for gene hunting

Workflow: NGS study



Design experiment

Library prep

Sequencing

Data analysis

Biological implication

Biological insights

Experimental design

- Issues to be considered in designing RNA-seq experiments.
  - You should define the **goal**.
  - Which **platform** do you choose?
  - **Depth**: How many reads do you need per sample?
  - **Length**: How long do you sequence?
  - **Paired-end** or single-end?
  - Method for **library construction**
    - Strand-specific?
    - Normalize?
  - How many biological **replicates**?
  - Pool RNA from multiple individuals or use a single individual?
  - Batch effect and lane effect.
  - **Informatics** strategy.

Experimental design for **gene cataloguing**

- **Depth**: How many reads do you need per sample?
- **Length**: How long do you sequence?
- **Paired-end** or single-end?
- Method for **library construction**
  - Strand-specific?
  - Normalize?
- How many biological **replicates**?
- Pool RNA from multiple?
- **Informatics** strategy.

- Difficult question…

- Longer is better.
- Paired-end is strongly recommended.

(ex) PE:100+100

- Strand-specific library is preferred, but normal one works well enough.
- Normalized library is not recommended.

- No replicates required. Instead
- Collect RNA from a wide variety of samples: tissue, cell type, developing stage (age), sex, treatments, environment etc.
- Single individual is preferred

Experimental design for **DE analysis**

- **Depth**: How many reads do you need per sample?
- **Length**: How long do you sequence?
- **Paired-end** or single-end?
- Method for **library construction**
  - Strand-specific?
  - Normalize?
- How many biological **replicates**?
- Pool RNA from multiple?
- **Informatics** strategy.

- Difficult question…

- If you have reference, single-end shorter reads are good enough. (ex. SE: 50 ~ 75)

- Normal TruSeq is good enough for most purposes.
- Consider strand-specific library if you want to know anti-sense RNA etc.

- Biological replicates are strongly recommended.

Take-home message

RNA-seq is the powerful tool for studies of non-model organisms. It can produce a nearly complete picture of transcriptomic events in a biological sample.

DNA binding domain
RECOGNITION SITE
RECOGNITION SITE
TAL Effector
DNA binding domain

NGS

Genome Editing

Imaging

Acknowledgements

基生研・生物機能情報分析室のメンバー
北大・三浦研究室
基生研共同利用研究の研究者の皆様
植物最先端研究拠点ネットワーク
イルミナ

歓迎！大学院生（総研大）、ポスドク

Every organism that *excites* you is your MODEL