

# RNA-seq解析パイプライン上級： *de novo* RNA-seq, single-cell RNA-seq

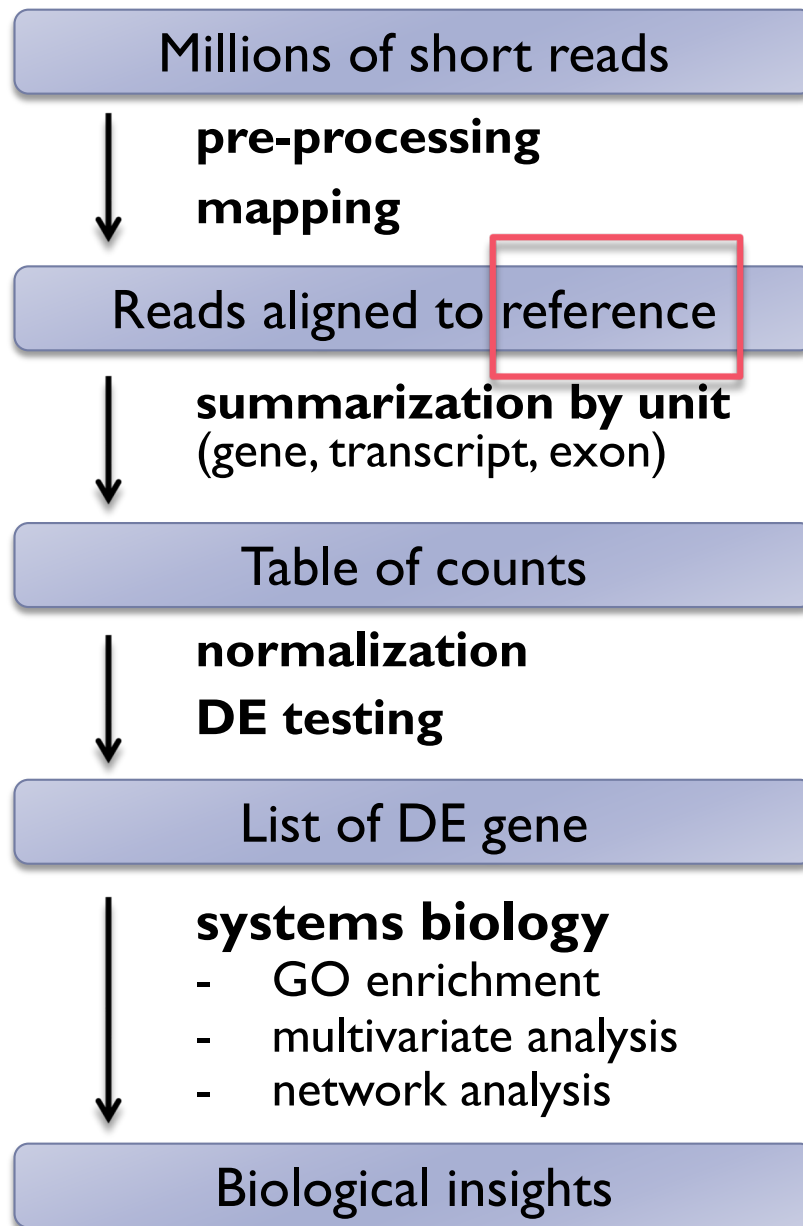
Shuji Shigenobu

重信 秀治

基礎生物学研究所  
生物機能解析センター

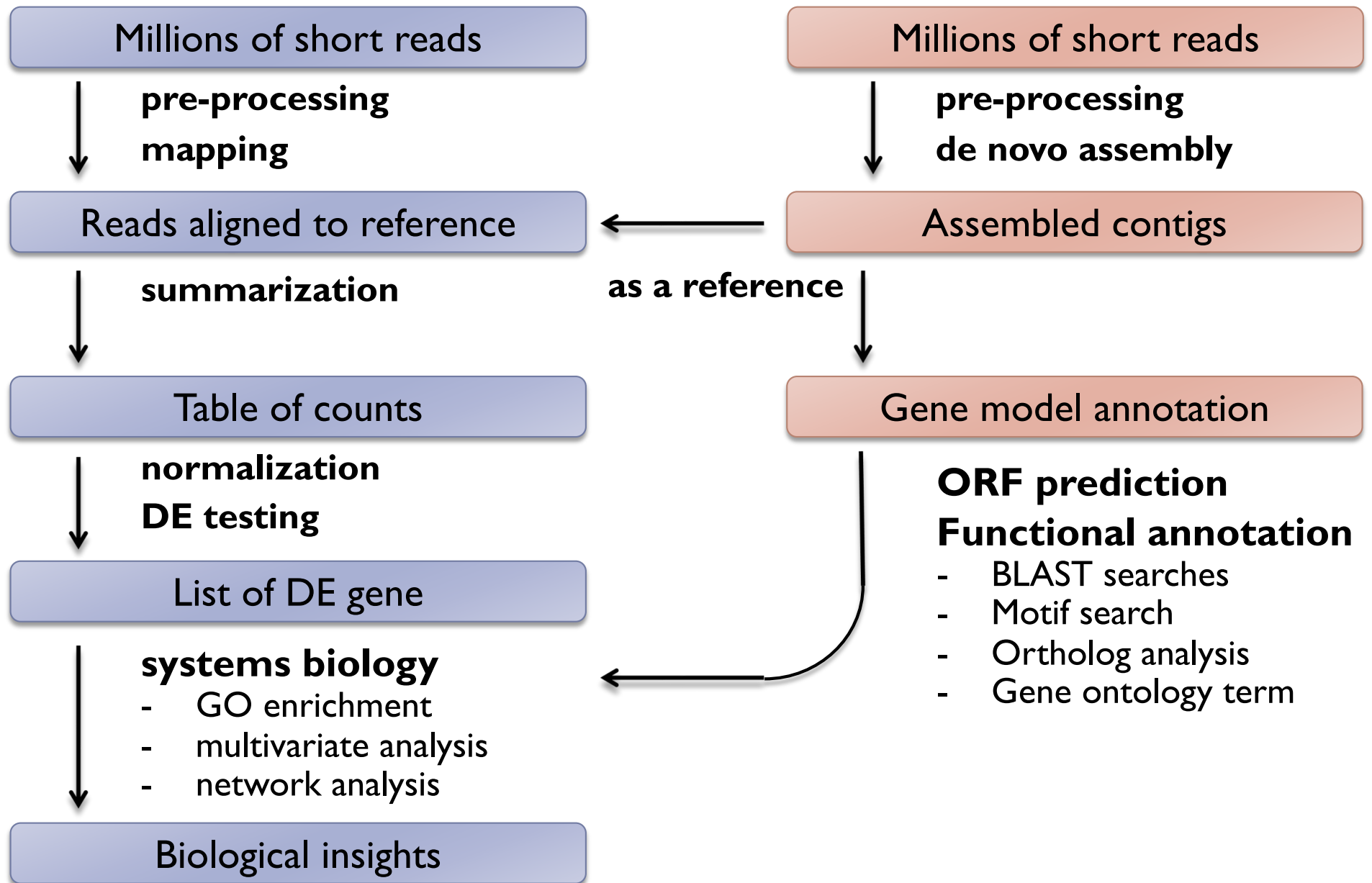


*de novo* RNA-seq



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

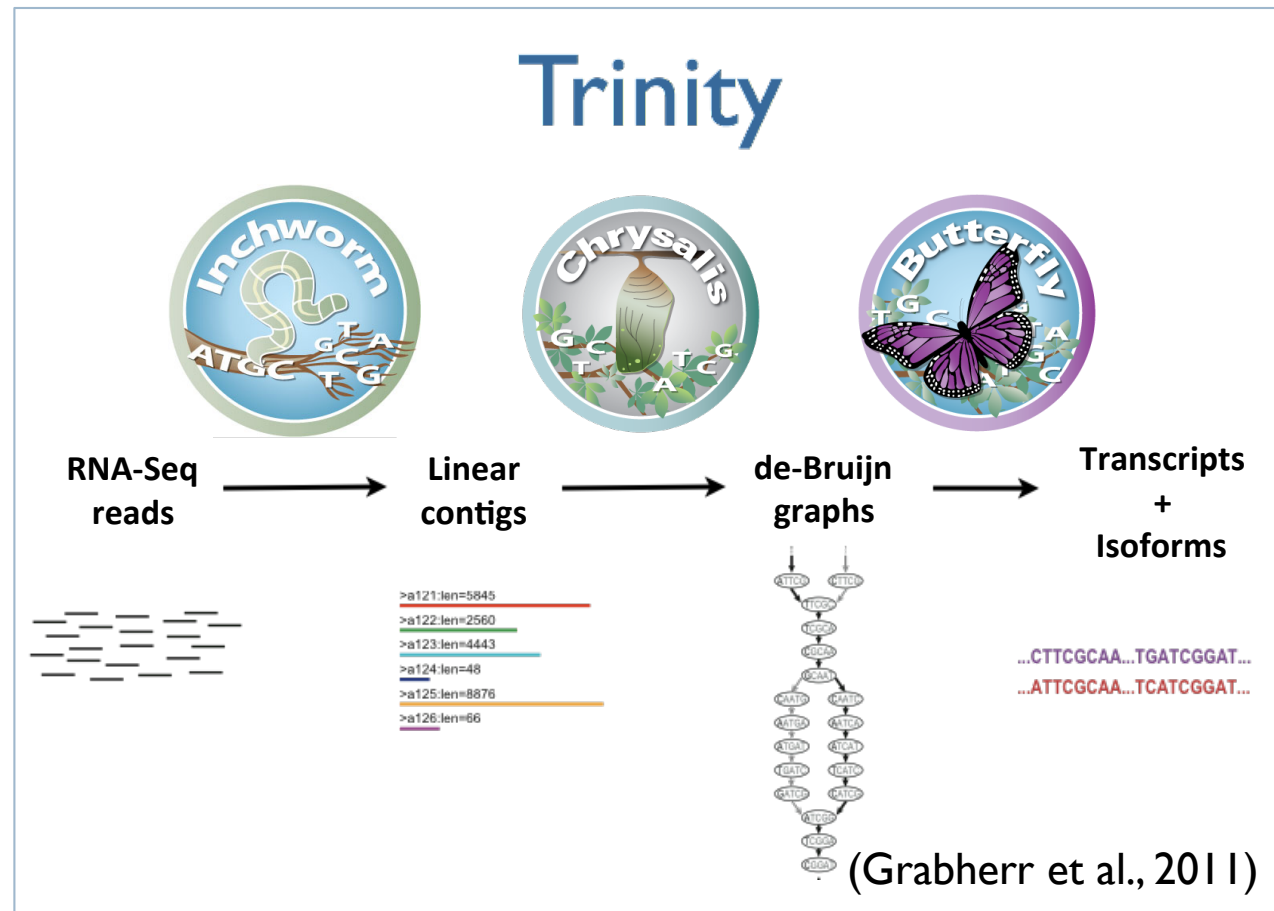
# RNA-seq analysis pipeline (*de novo* strategy)



# *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

# BUSCO

<https://busco.ezlab.org/>

**BUSCO**



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)

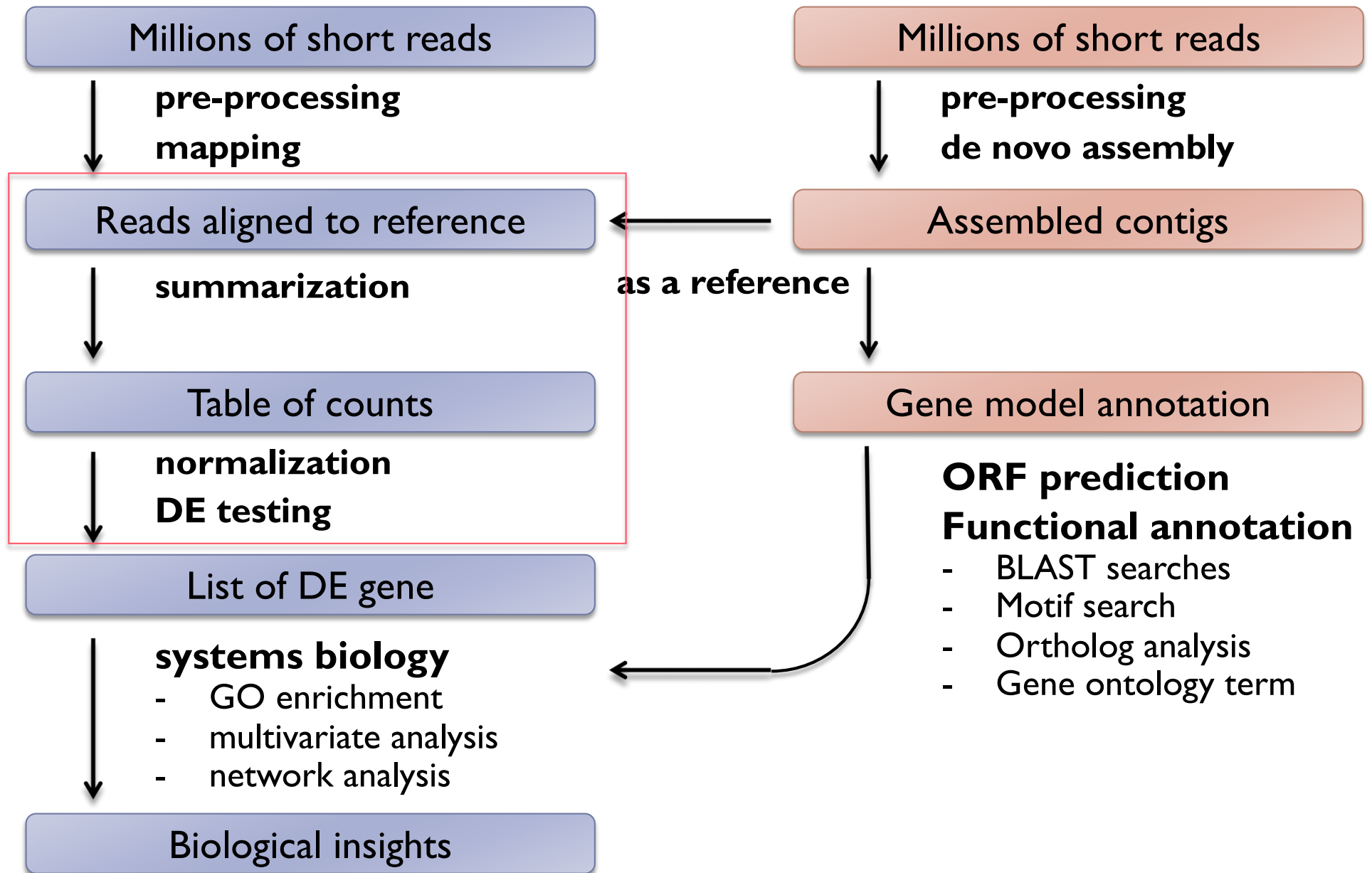
## ► Clustering

- 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.

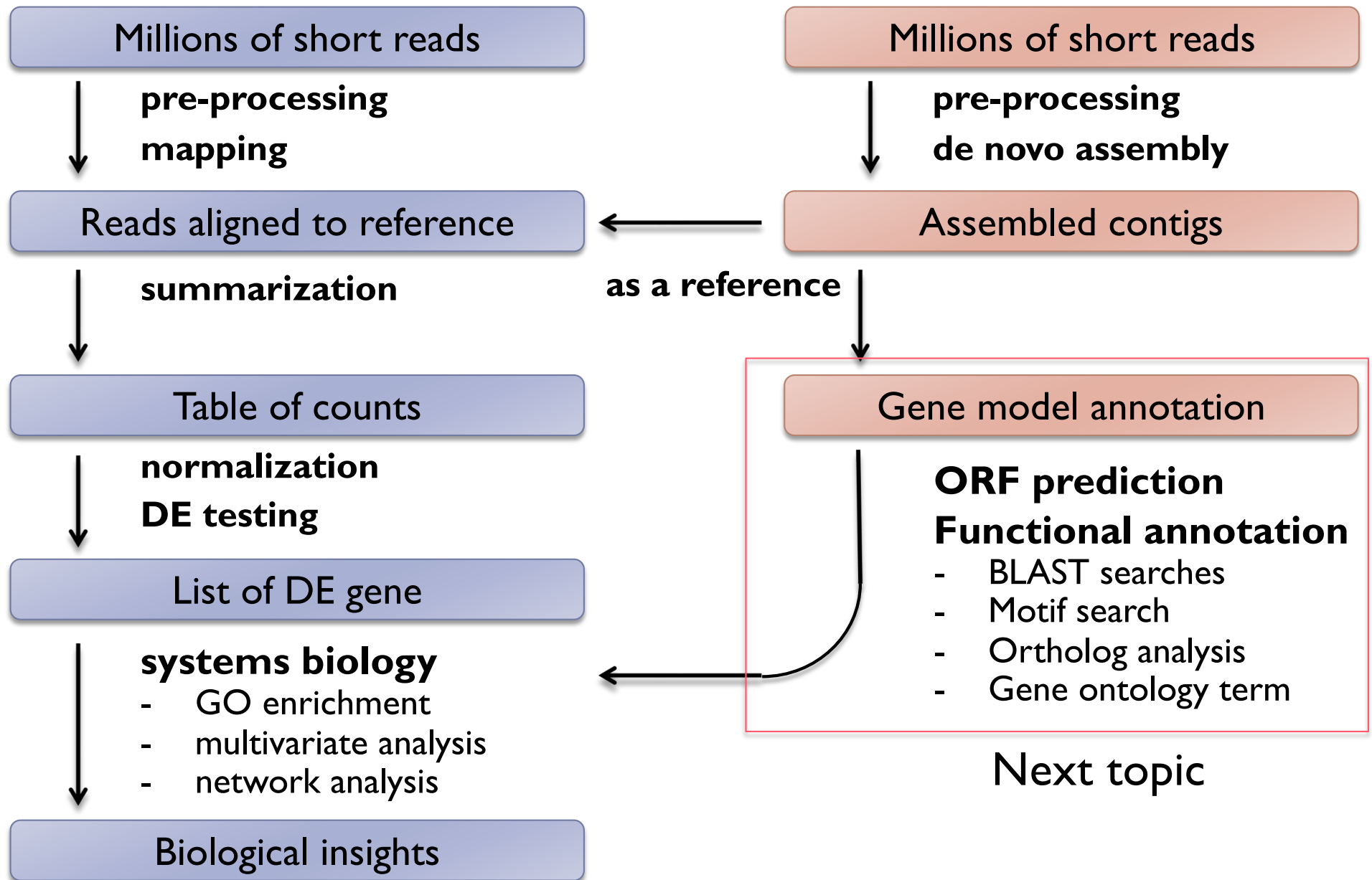
# RNA-seq analysis pipeline (*de novo* strategy)



# DEG analysis

- ▶ Follow transcript-based RNA-seq pipeline

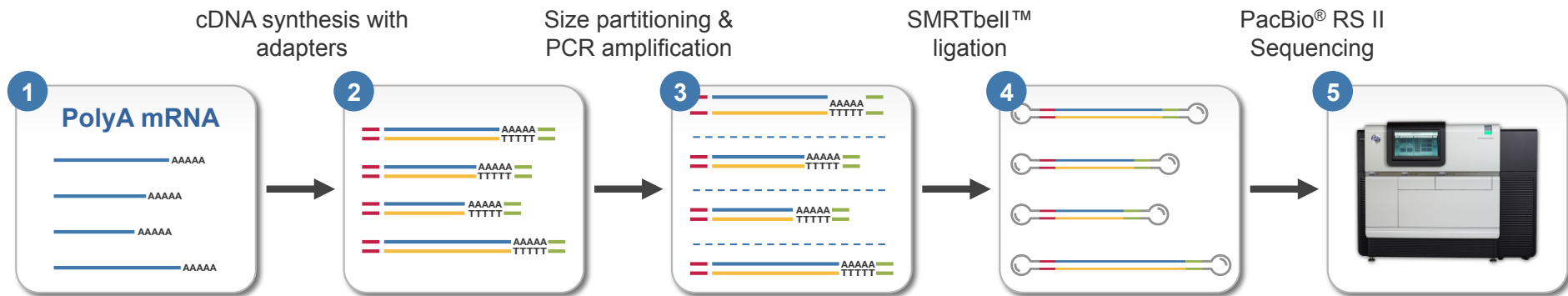
# RNA-seq analysis pipeline (*de novo* strategy)



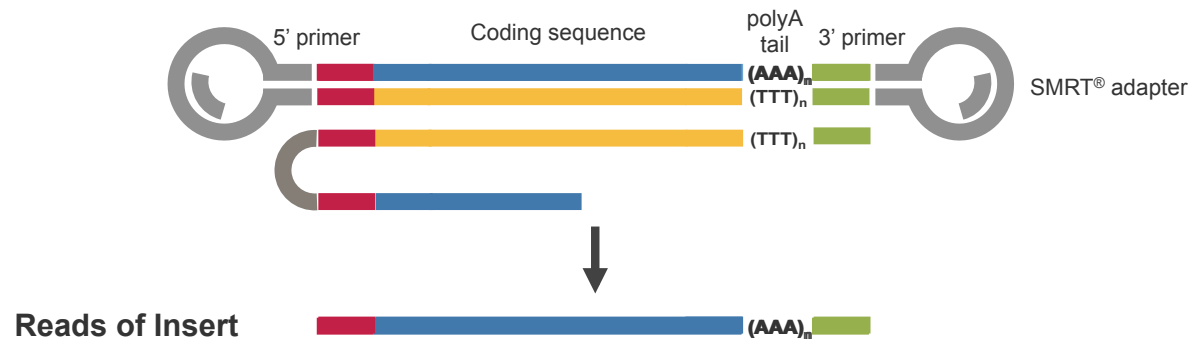


# PacBio Iso-Seq for building a transcriptome catalogues

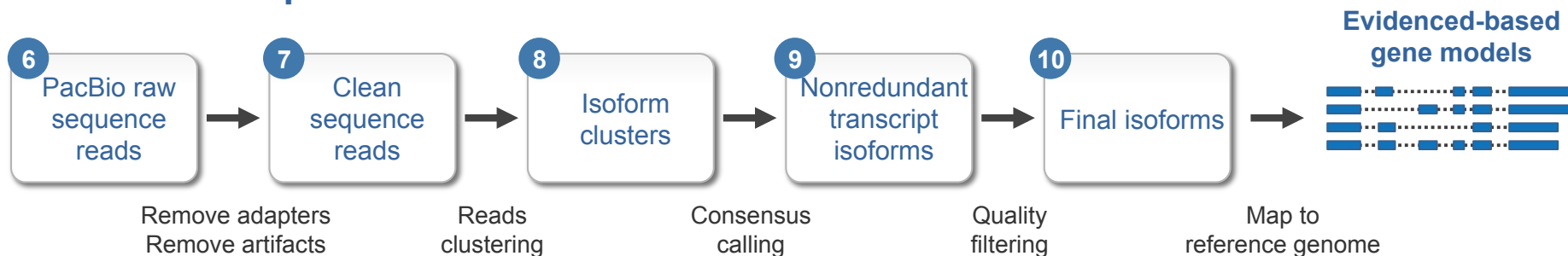
## Experimental Pipeline



## [SampleNet: Iso-Seq Method with Clontech® cDNA Synthesis Kit](#)

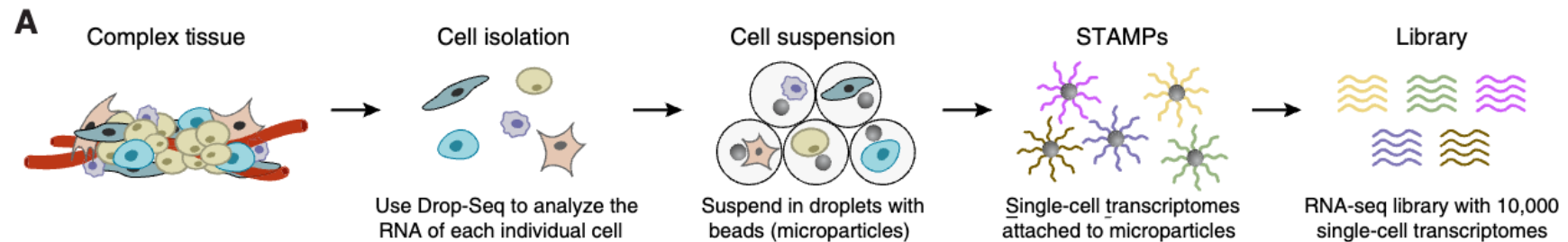


## Informatics Pipeline



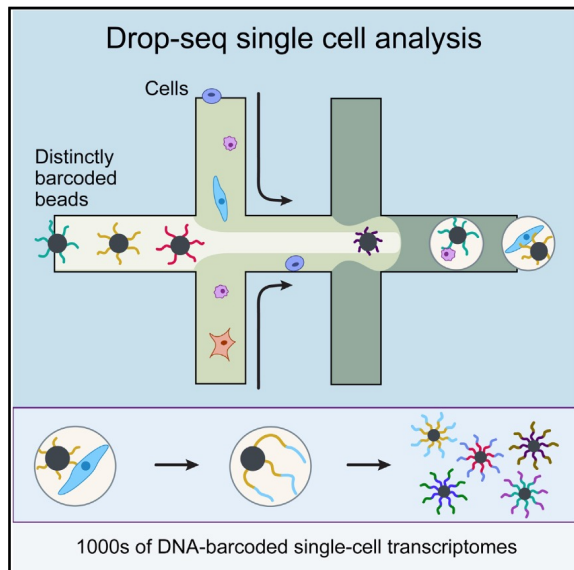
# Single-cell RNA-seq

# Drop-seq / Single-cell RNA-seq



## Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets

### Graphical Abstract



### Authors

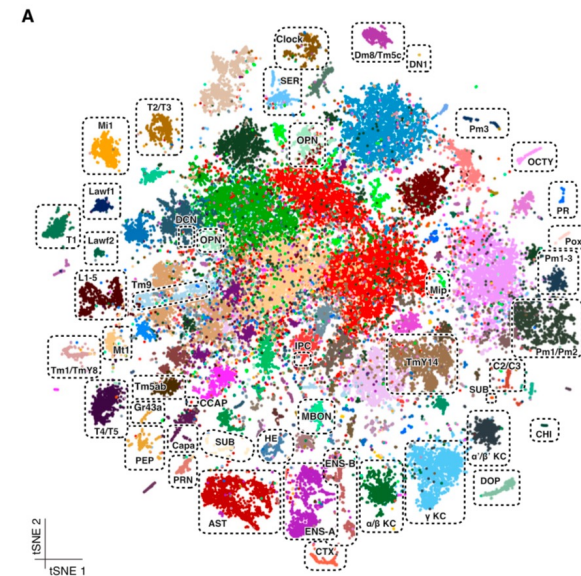
Evan Z. Macosko, Anindita Basu, ...,  
Aviv Regev, Steven A. McCarroll

### Correspondence

emacosko@genetics.med.harvard.edu  
(E.Z.M.),  
mccarroll@genetics.med.harvard.edu  
(S.A.M.)

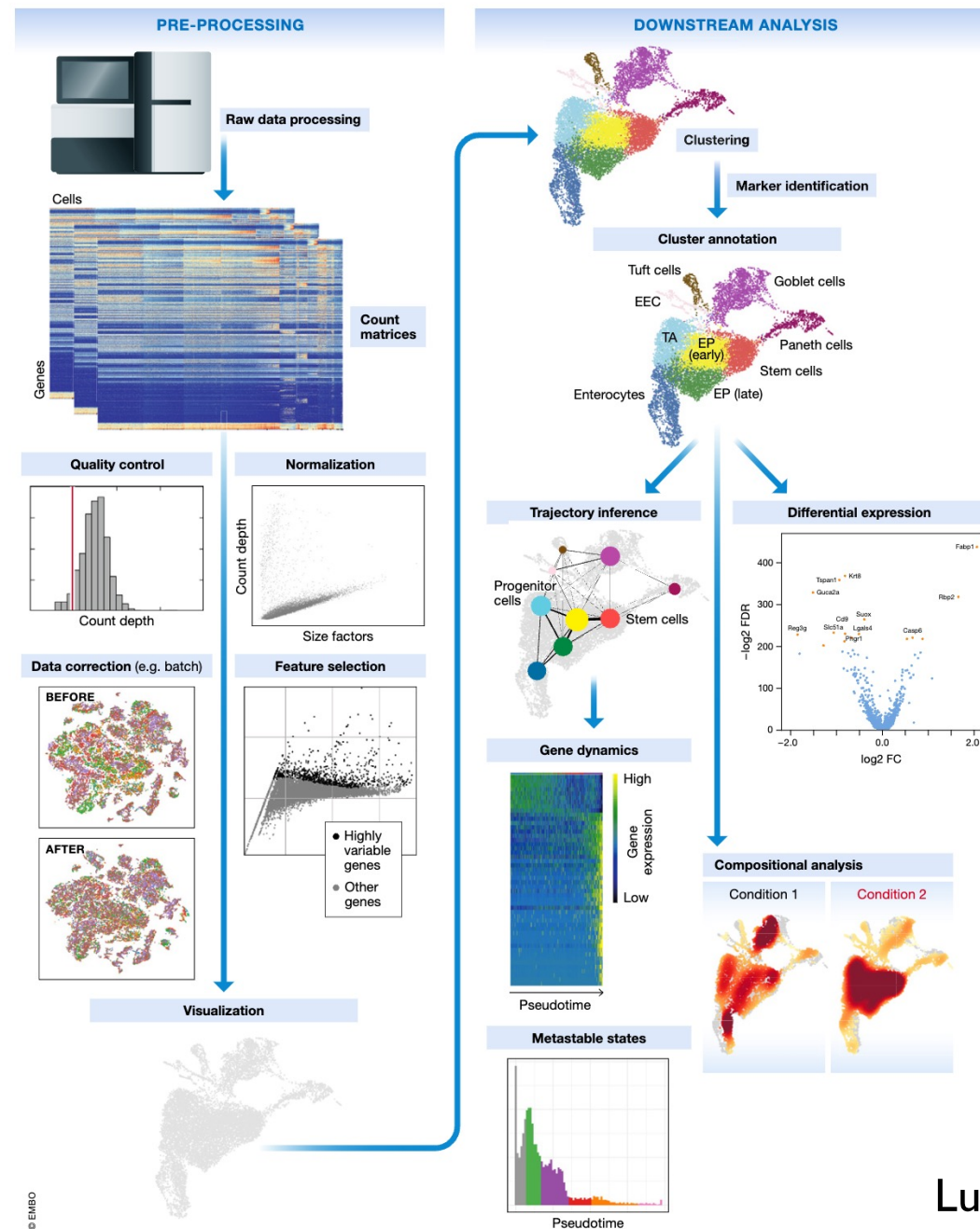
### In Brief

Capturing single cells along with sets of uniquely barcoded primer beads together in tiny droplets enables large-scale, highly parallel single-cell transcriptomics. Applying this analysis to cells in mouse retinal tissue revealed transcriptionally distinct cell populations along with molecular markers of each type.



(Macosko et al., 2015)

# Typical single-cell RNA-seq bioinformatics workflow



Luecken et al., 2022

# Bioinformatic of single-RNA-seq

- ▶ 代表的なプラットフォーム 10x Genomics Chromium. 数千細胞の transcriptome。
- ▶ 観測細胞が多いだけで、genes x cells のカウントマトリックスを扱う点は、Bulk RNA-seq と同じ。したがってバイオインフォマティクスの基礎は同じ。
- ▶ とはいえ、RNA-seq特有の問題も多く、RNA-seqに特化したアルゴリズム・ソフトウェアが活発に開発されている。
- ▶ Bulk RNA-seq と異なる点
  - ▶ Sparse data (ゼロカウントの遺伝子が多い) 。それゆえ、データはnoisy。
  - ▶ 観測細胞が桁違いに多い
  - ▶ 細胞のクラスタリングに重きを置いた解析が多い
  - ▶ scRNA-seqならではの解析の例として、psudotime 解析など
- ▶ Popular tools
  - ▶ CellRanger: 10x Genomics社純正 QC + mapping + count matrix generation
  - ▶ Seurat: integrated analysis platform (from QC to clustering)

# Sparse matrix data

- ▶ scRNA-seq data matrix is “sparse” matrix (many zero count)
- ▶ Rather than the regular CSV format, sparse formats (only the nonzero entries are stored) are preferred.
- ▶ CellRanger uses Market Exchange Format (MEX)

# MEX

```
$ tree filtered_feature_bc_matrix
```

```
filtered_feature_bc_matrix
```

```
├── barcodes.tsv.gz
```

```
├── features.tsv.gz
```

```
└── matrix.mtx.gz
```

```
[features.tsv]
```

```
ENSG000000141510
```

```
TP53
```

```
Gene Expression
```

```
ENSG000000012048
```

```
BRCA1
```

```
Gene Expression
```

```
ENSG000000139687
```

```
RB1
```

```
Gene Expression
```

```
CD3_GCCTGACTAGATCCA
```

```
CD3
```

```
Antibody Capture
```

```
CD19_CGTGCAACACTCGTA
```

```
CD19
```

```
Antibody Capture
```

```
[barcodes.tsv]
```

```
AAACCCAAGGAGAGTA-1
```

```
AAACGCTTCAGCCCAG-1
```

```
AAAGAACAGACGACTG-1
```

```
AAAGAACCAATGGCAG-1
```

```
[matrix.mtx]
```

```
%%MatrixMarket matrix coordinate real general
```

```
%
```

```
32738 2700 2286884
```

```
32709 1 4
```

```
32707 1 1
```

```
32706 1 10
```

```
32704 1 1
```

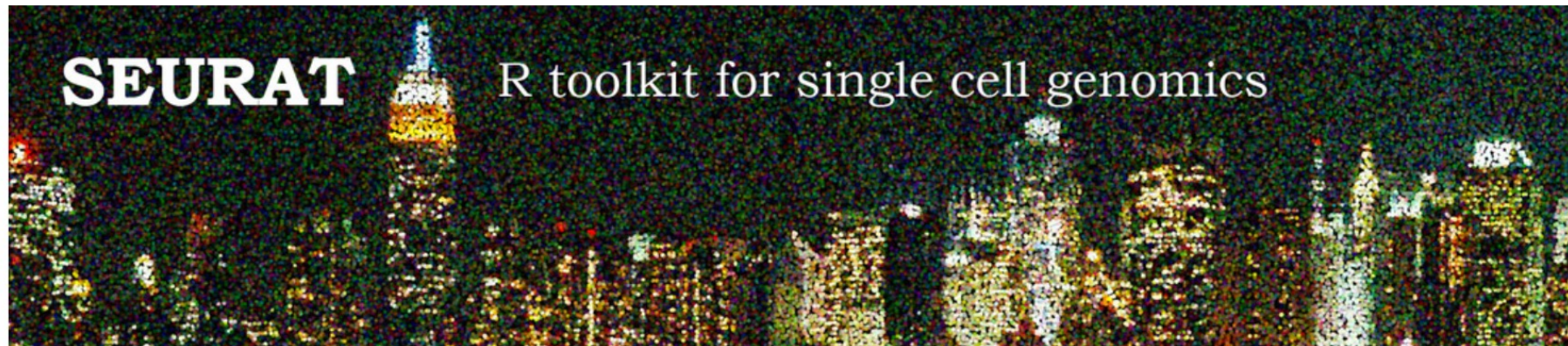
```
32703 1 5
```



# Seurat

<https://satijalab.org/seurat/>

Seurat 4.0.6 Install Get started Vignettes ▾ Extensions FAQ News Reference Archive



- Seurat is an R package designed for QC, analysis, and exploration of single-cell RNA-seq data.
- Nice documentation including detailed tutorials and FAQ.
- Frequent updates incorporating new algorithms.