

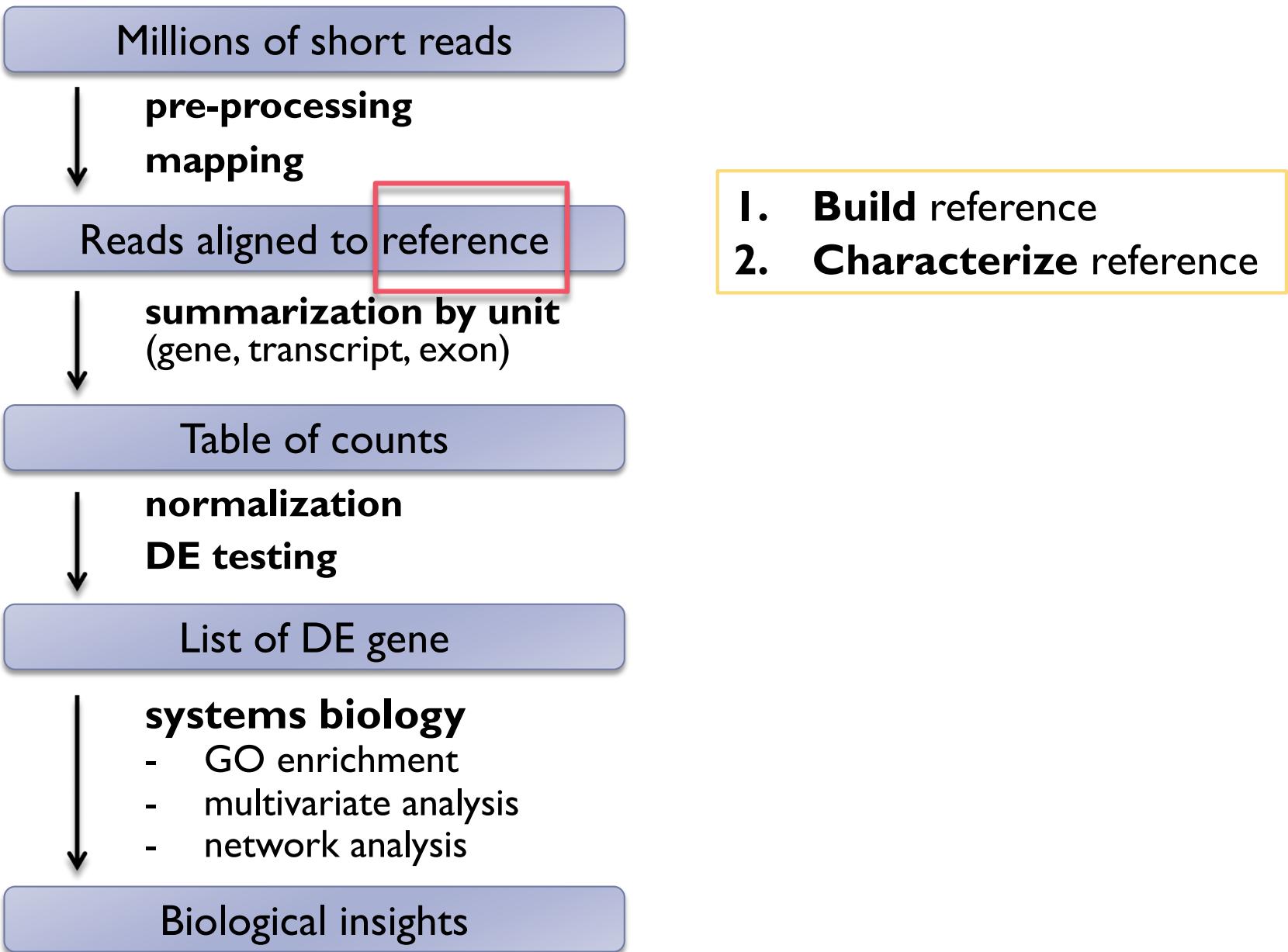
RNA-seq解析パイプライン上級： *de novo assembly &* single-cell RNA-seq

Shuji Shigenobu
重信 秀治

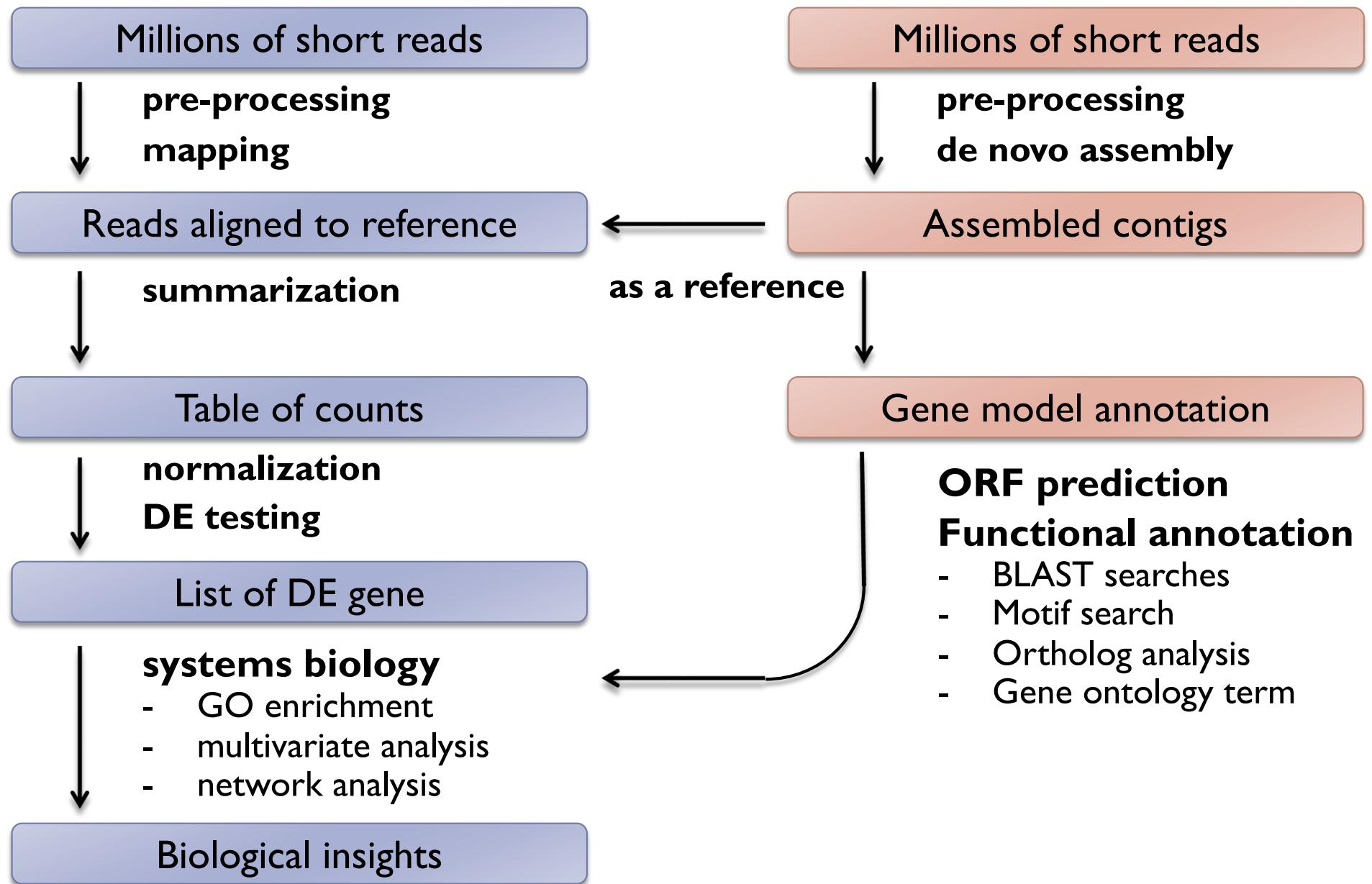
基礎生物学研究所



de novo RNA-seq assembly



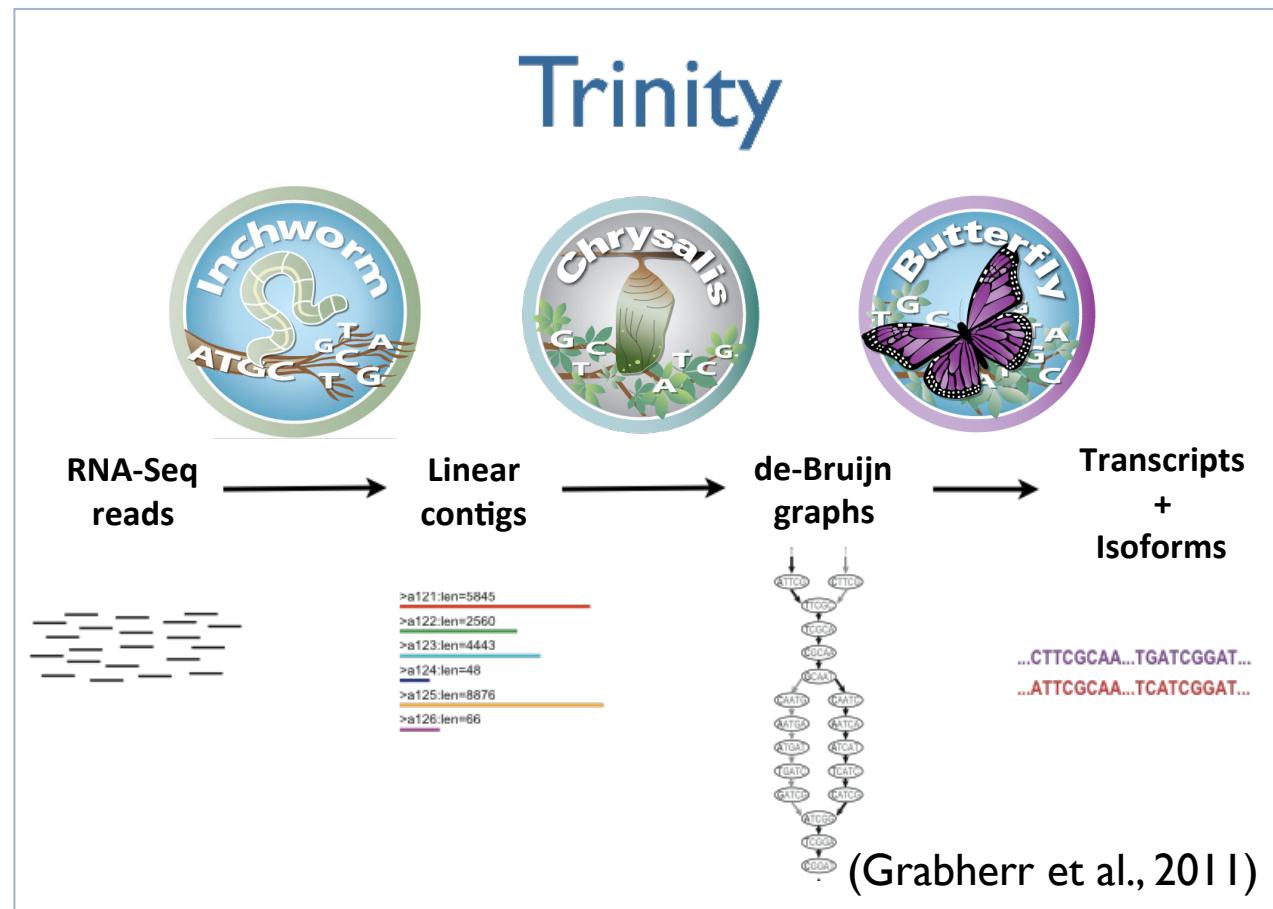
RNA-seq analysis pipeline (*de novo* strategy)



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
- ▶ ...



[Code](#)[Issues 25](#)[Pull requests 0](#)[Projects 0](#)[Wiki](#)[Insights](#)

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

► 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

demo

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

Typical assembly stats

```
#####
## Counts of transcripts, etc.
#####

Total trinity 'genes': 7092
Total trinity transcripts: 7269
Percent GC: 44.84

#####
Stats based on ALL transcript contigs:
#####

Contig N10: 1255
Contig N20: 914
Contig N30: 714
Contig N40: 564
Contig N50: 453
Median contig length: 301
Average contig: 414.20
Total assembled bases: 3010808
```

N50

- ▶ Definition: Scaffold/contig length at which you have covered 50% of total assembly length
- ▶ Larger N50 means longer contiguity

Example and how to calculate

You have 5 contigs:

- 200, 800, 100, 1500, 1000 bp

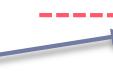
Calculate total length => 3600 bp

Sort by length and calculate cumulative length

Find the contig at which cumulative / total length reach 50%

Contig size	Cumulative (bp)	Cumulative / total length
1500	1500	41.6%
1000	2500	69.4%
800	3300	91.6%
200	3500	97.2%
100	3600	100%

N50



50%

N50

- ▶ N50の求め方
 - ▶ seqkit ソフトウェアなど
 - ▶ 定義は簡単なので自分でコードを書く

(R)

```
> len <- c(200, 800, 100, 1500, 1000)
> len.sorted <- rev(sort(len))
> N50 <- len.sorted[cumsum(len.sorted) >= sum(len.sorted)*0.5][1]
```



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

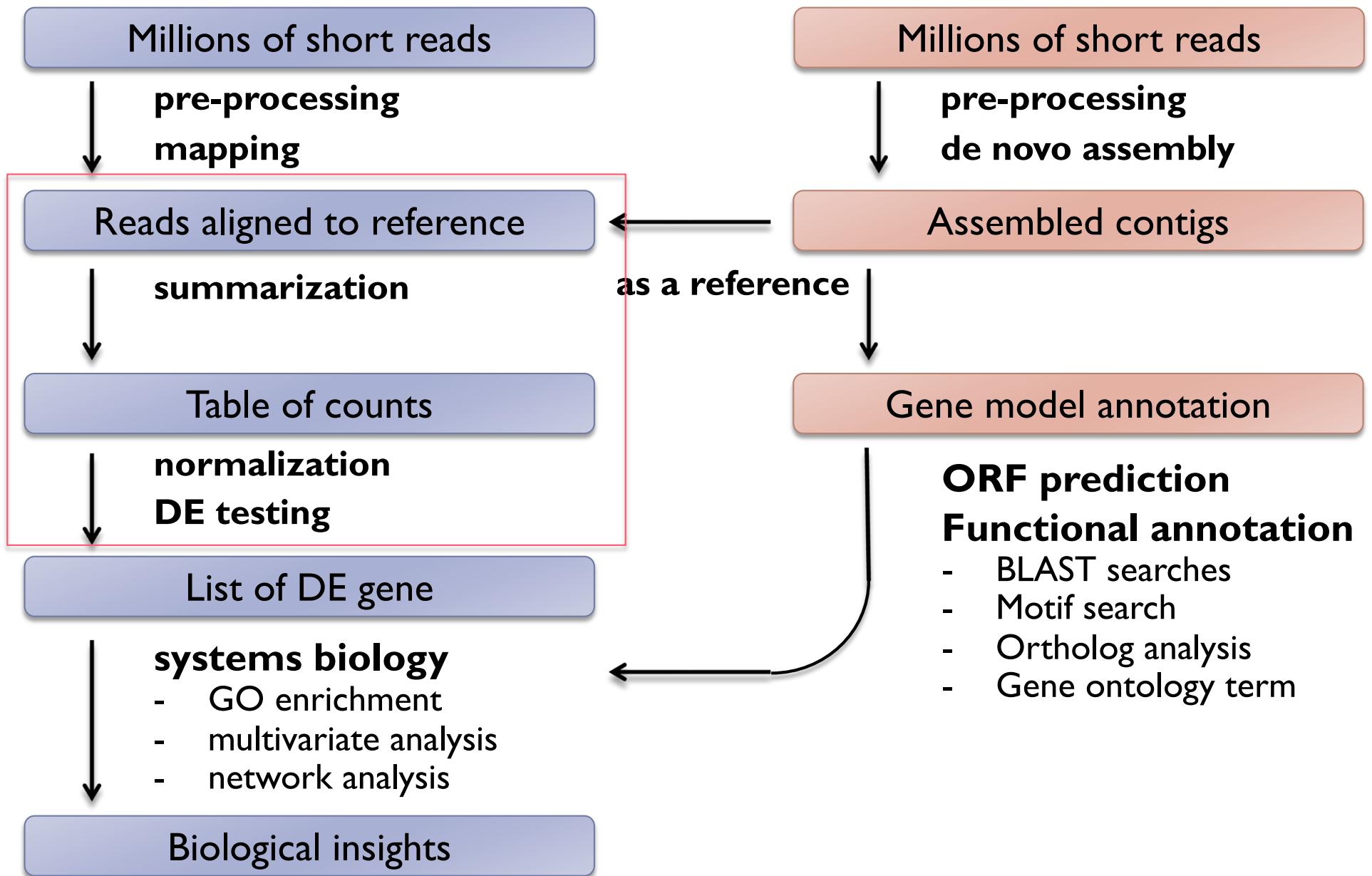
- ▶ Choose one representative transcript from each cluster based on Trinity component information. (longest or highest expression)
- ▶ Clustering

- ▶ CDHIT-EST (<http://weizhongli-lab.org/cd-hit/>)
- ▶ Corset (Davidson et al., 2014).
- ▶ RapClust (<https://github.com/COMBINE-lab/RapClust>)
- ▶ EvidentialGene
(<http://arthropods.eugenesc.org/EvidentialGene/trassembly.html>)

▶ Advantage of redundancy reduction

- ▶ Gene-oriented analysis => easier interpretation
- ▶ Better control of multiple comparison.

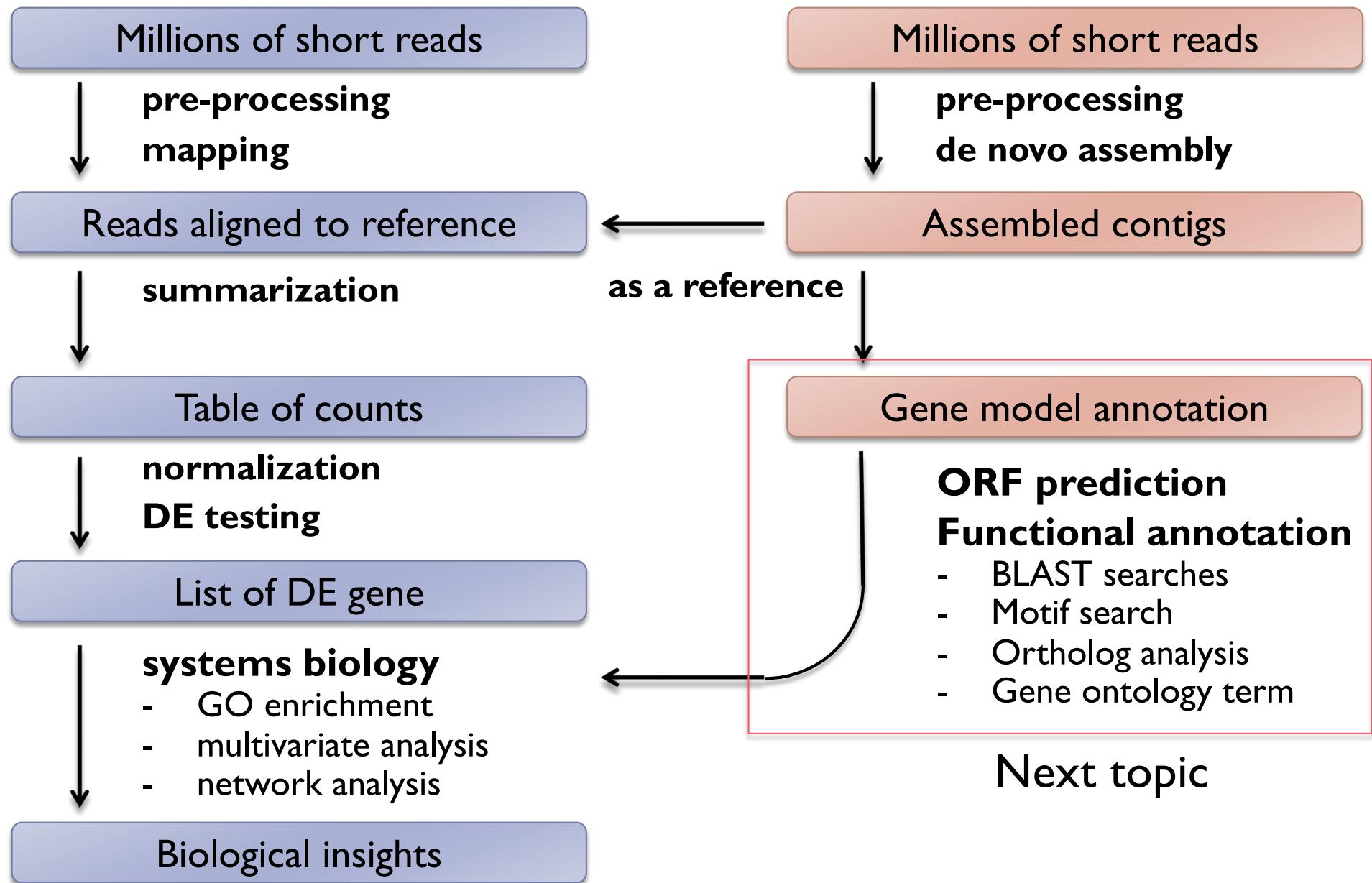
RNA-seq analysis pipeline (*de novo* strategy)



DEG analysis

- ▶ Follow transcript-based RNA-seq pipeline

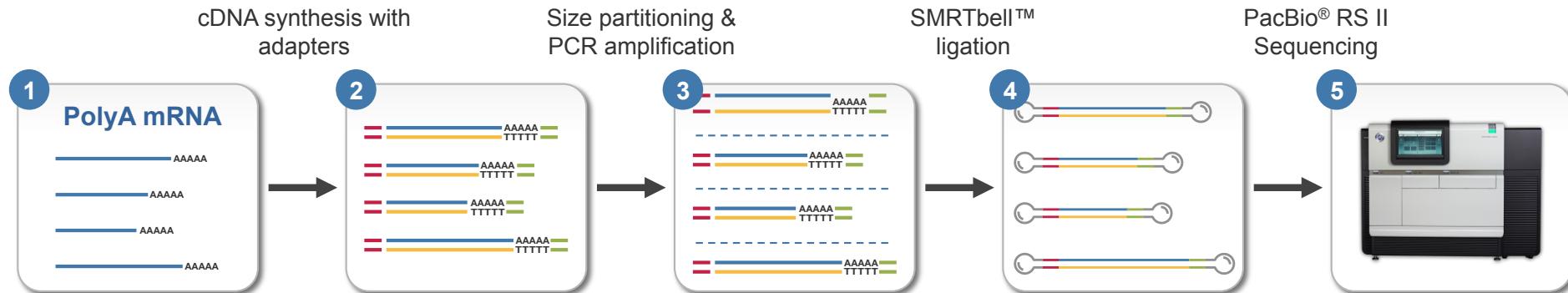
RNA-seq analysis pipeline (*de novo* strategy)



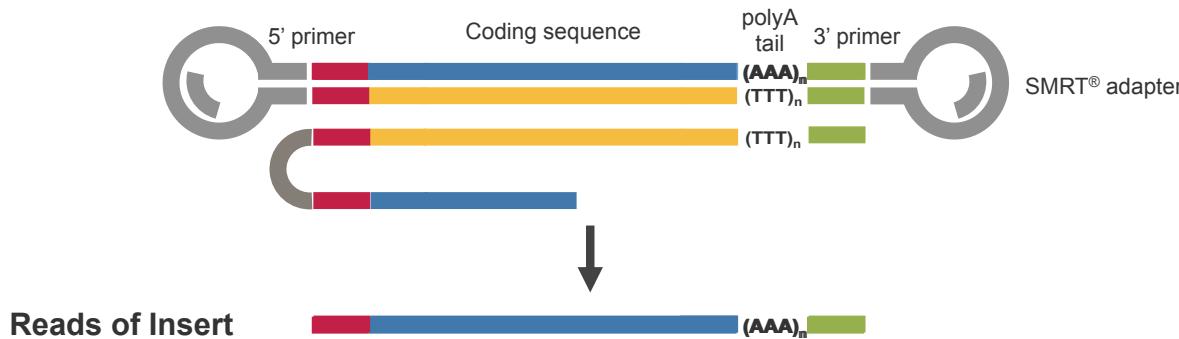
PacBio Iso-Seq for building a transcriptome catalogues

Now Kinnex is available

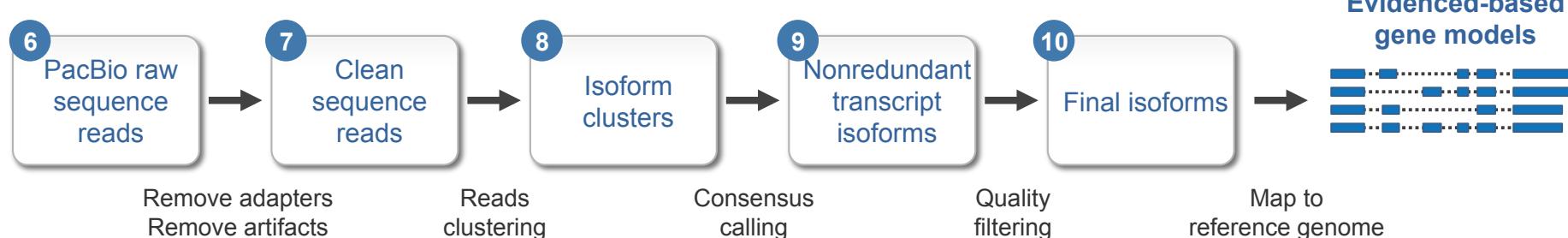
Experimental Pipeline



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

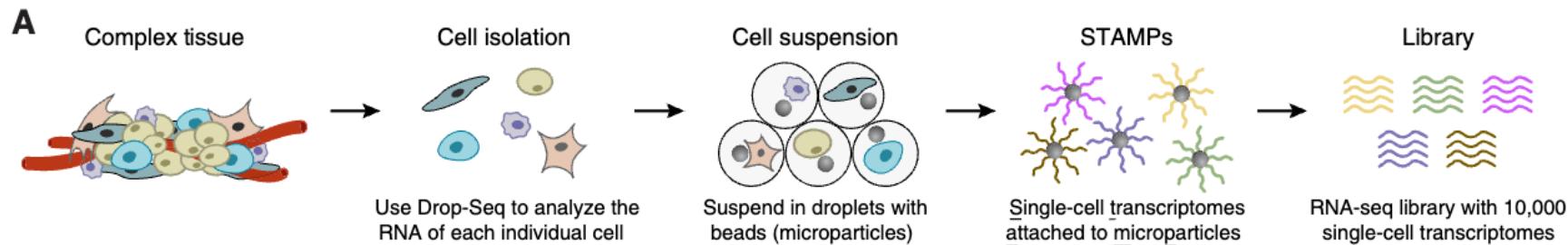


Informatics Pipeline



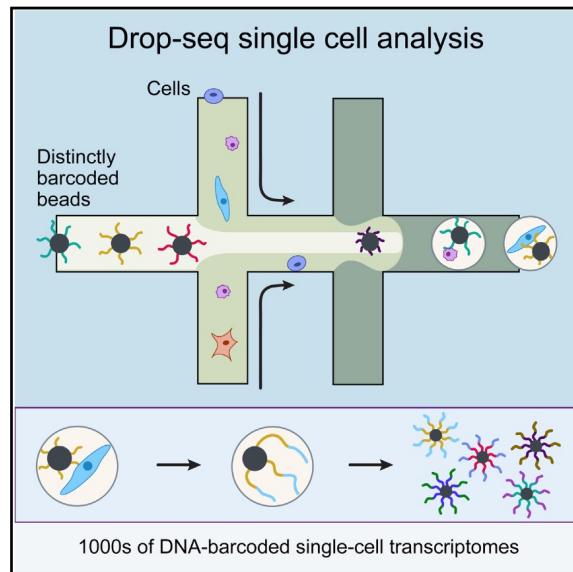
Single-cell RNA-seq

Drop-seq / Single-cell RNA-seq



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

Graphical Abstract



Authors

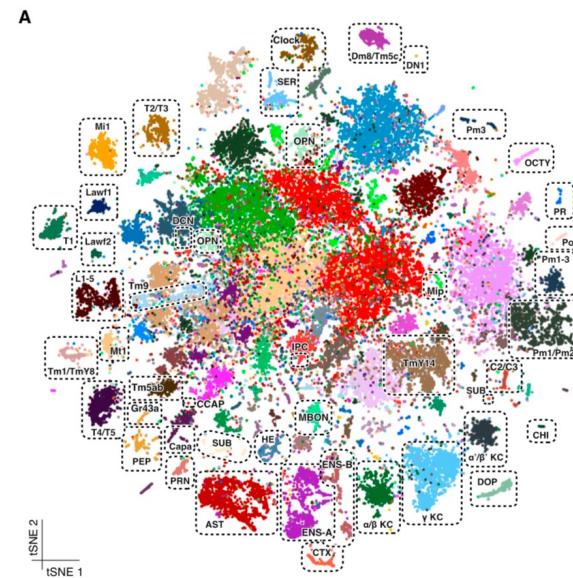
Evan Z. Macosko, Anindita Basu, ...
Aviv Regev, Steven A. McCarroll

Correspondence

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

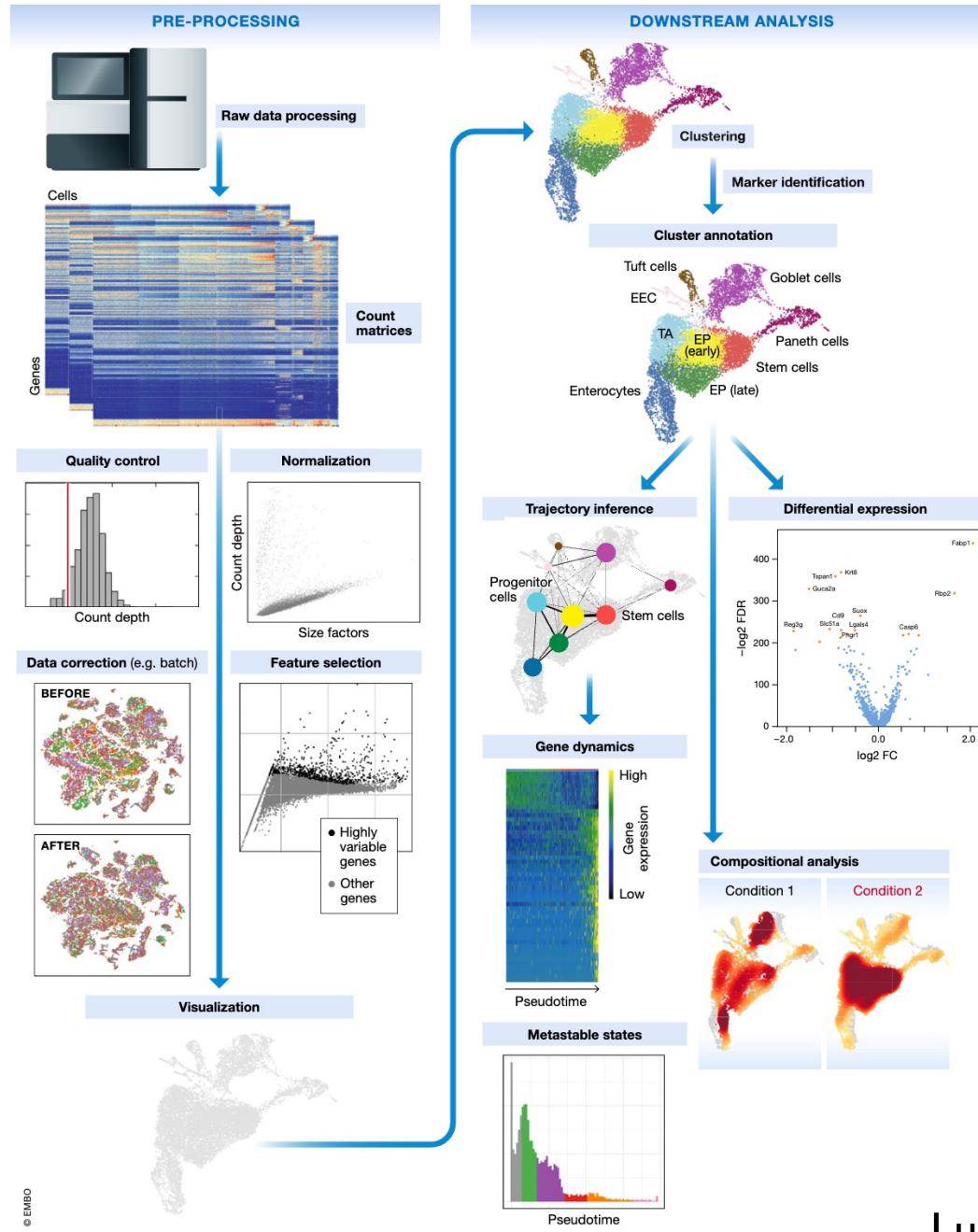
In Brief

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

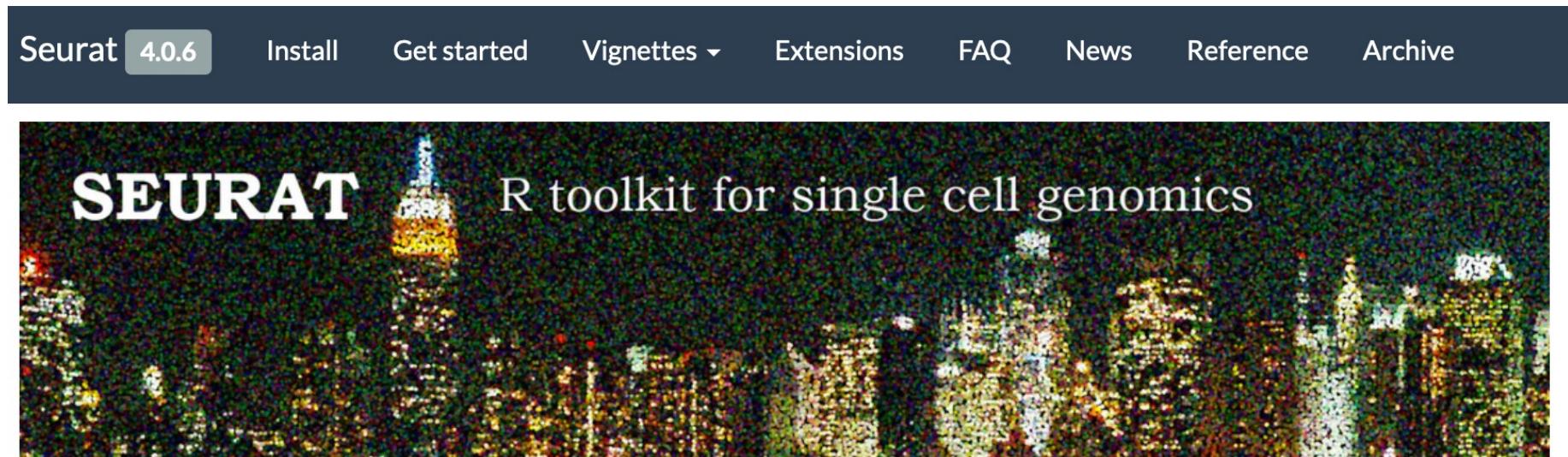
Bioinformatics of single-RNA-seq

- ▶ 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 × 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、scRNA-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
 - ▶ Sparce data (ゼロカウントの遺伝子が多い)。それゆえ、データはnoisy。
 - ▶ UMIを導入しているプラットフォームでは、生のリードカウントではなく UMIを使う。
 - ▶ 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。pseudotime 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)
- ▶ Geneformerなど、deep learning モデルへの期待

Seurat

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

R



R



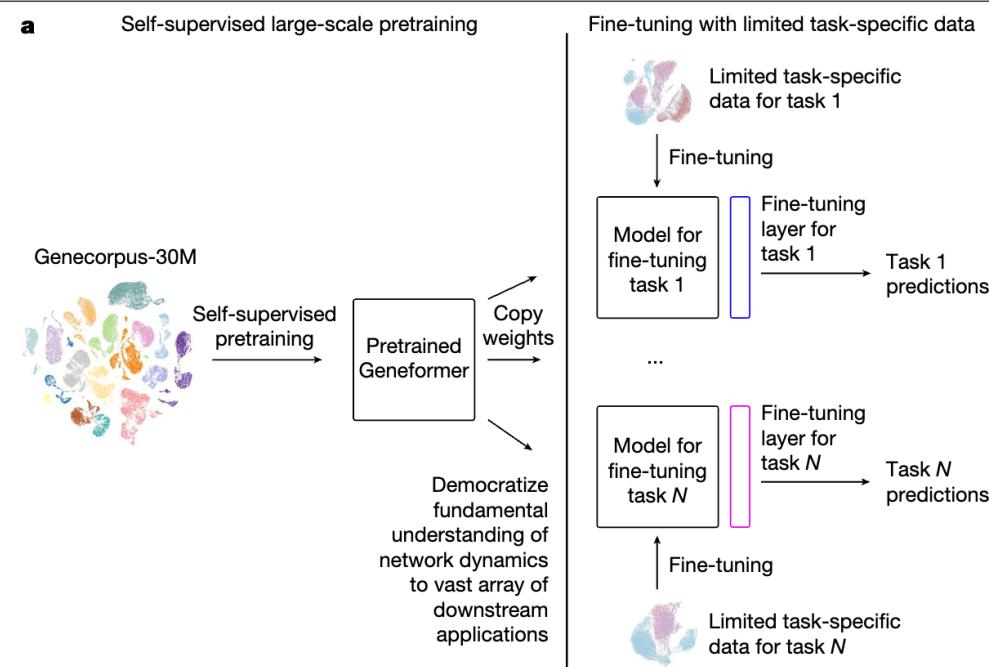
Python

From scRNA-seq to AI

- ▶ **Geneformer**
- ▶ **scGPT**
- ▶ ...

Article

Transfer learning enables predictions in network biology



Geneformer (Theodoris et al. Nature, 2023)