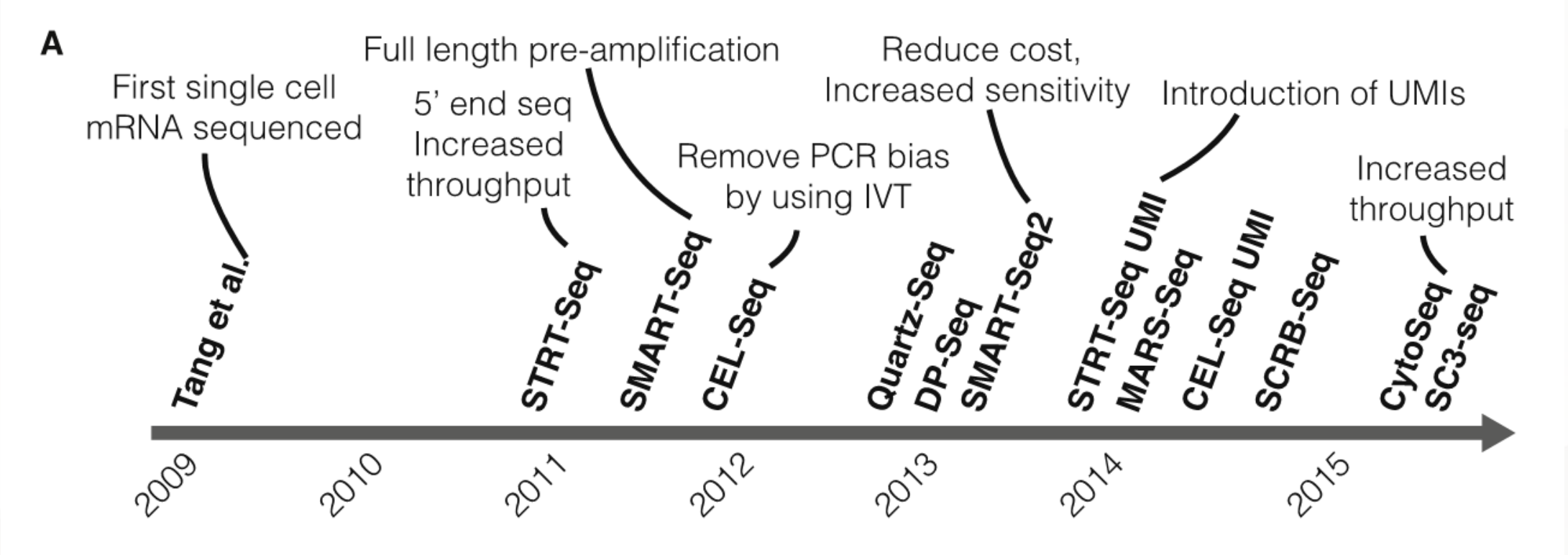
高性能计算-生命学科使用经验分享会

The development of scRNA-seq



sc-RNA technology acquired the 2018 break through of the year

Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis

BY JEFFREY A. FARRELL, YIQUN WANG, SAMANTHA J. RIESENFELD, KARTHIK SHEKHAR, AVIV REGEV, ALEXANDER F. SCHIER

01 JUN 2018

Single-cell RNA sequencing and a computational technique reveal cell trajectories that form the complex body plan of the zebrafish embryo.

ABSTRACT

The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution

BY JAMES A. BRIGGS, CALEB WEINREB, DANIEL E. WAGNER, SEAN MEGASON, LEONID PESHKIN, MARC W. KIRSCHNER, ALLON M. KLEIN

01 JUN 2018

A single-cell transcriptome analysis of whole frog embryos reveals cell states and provides a map of differentiation over time.

ABSTRACT

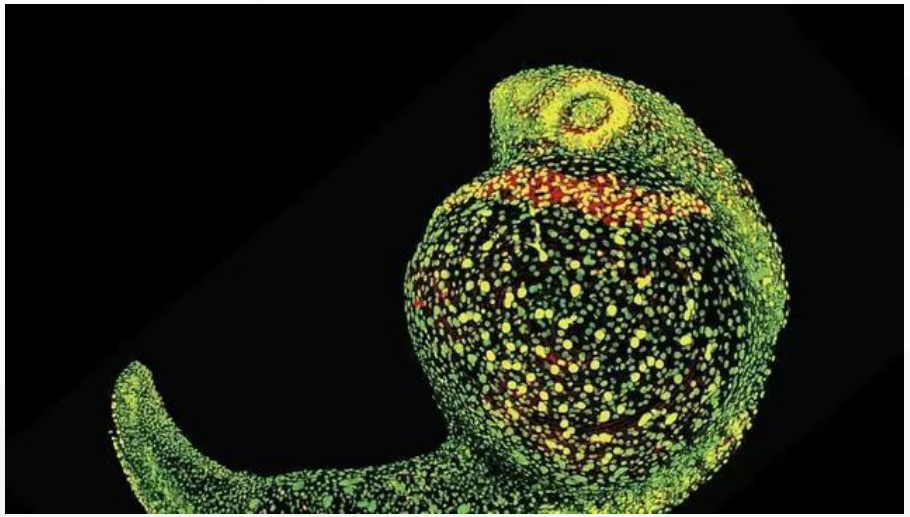
Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo

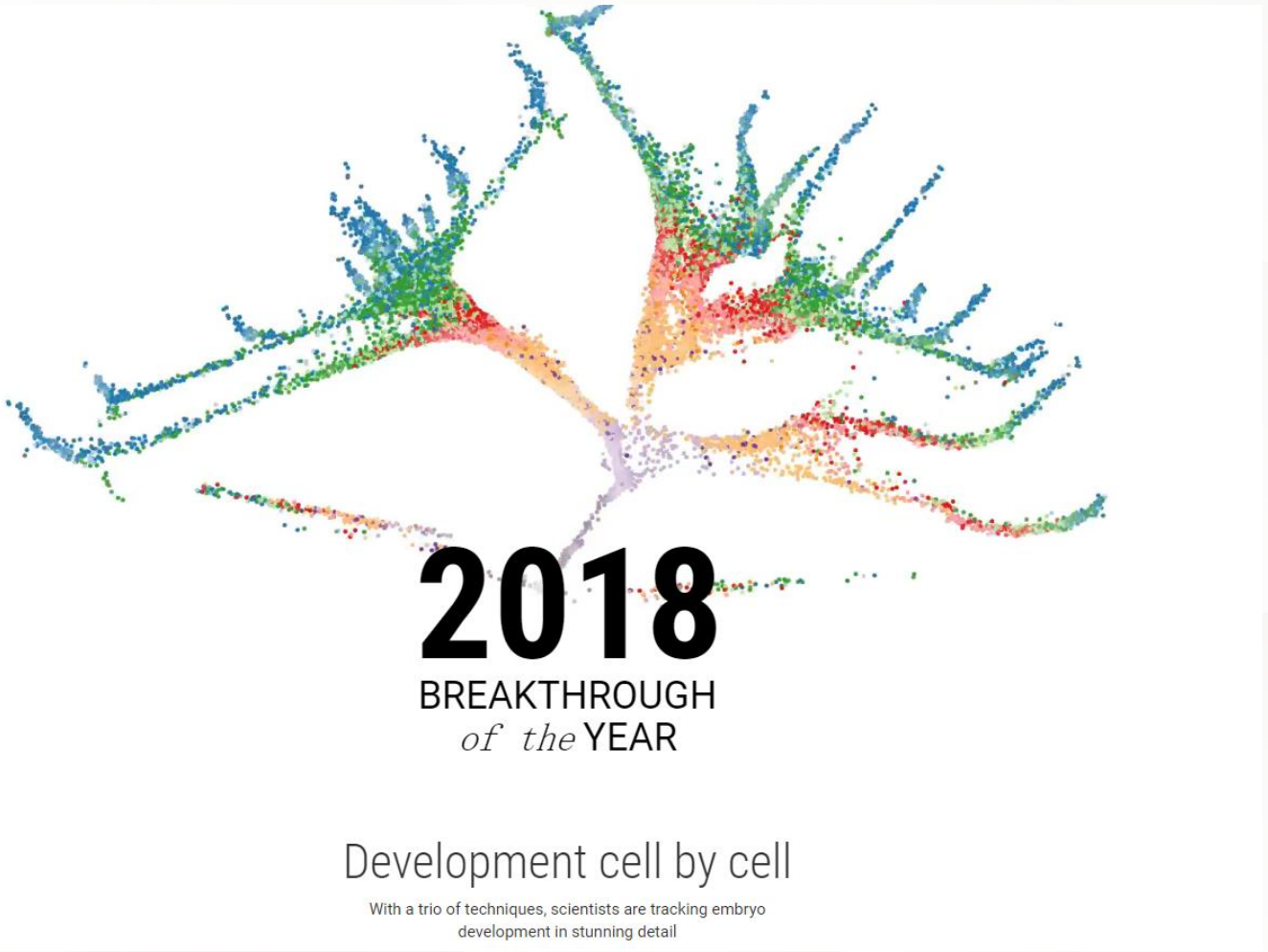
BY DANIEL E. WAGNER, CALEB WEINREB, ZACH M. COLLINS, JAMES A. BRIGGS, SEAN G. MEGASON, ALLON M. KLEIN

01 JUN 2018 : 981-987

Single-cell RNA sequencing reveals cell type trajectories and cell lineage in the developing zebrafish embryo.

ABSTRACT





science

上海科技大学
ShanghaiTech University

Common scRNA-seq platform

Smart-seq2

PROTOCOL

Full-length RNA-seq from single cells using Smart-seq2

Simone Picelli[1], Omid R Faridani[1], Åsa K Björklund[1,2], Gösta Winberg[1,3], Sven Sagasser[1,2] & Rickard Sandberg[1,2]

[1]Ludwig Institute for Cancer Research, Stockholm, Sweden. [2]Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden.
Correspondence should be addressed to R.S. (rickard.sandberg@ki.se).

Published online 2 January 2014; doi:10.1038/nprot.2014.006

Emerging methods for the accurate quantification of gene expression in individual cells hold promise for revealing the extent, function and origins of cell-to-cell variability. Different high-throughput methods for single-cell RNA-seq have been introduced that vary in coverage, sensitivity and multiplexing ability. We recently introduced Smart-seq for transcriptome analysis from single cells, and we subsequently optimized the method for improved sensitivity, accuracy and full-length coverage across transcripts. Here we present a detailed protocol for Smart-seq2 that allows the generation of full-length cDNA and sequencing libraries by using standard reagents. The entire protocol takes ~2 d from cell picking to having a final library ready for sequencing; sequencing will require an additional 1–3 d depending on the strategy and sequencer. The current limitations are the lack of strand specificity and the inability to detect nonpolyadenylated (polyA⁻) RNA.

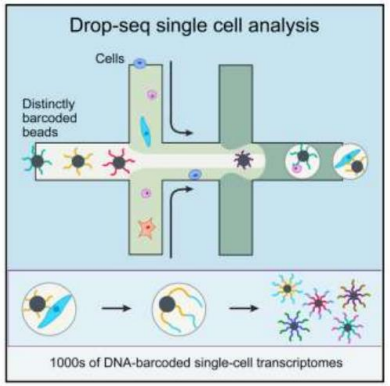
Drop-seq

Resource

Cell

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

Graphical Abstract

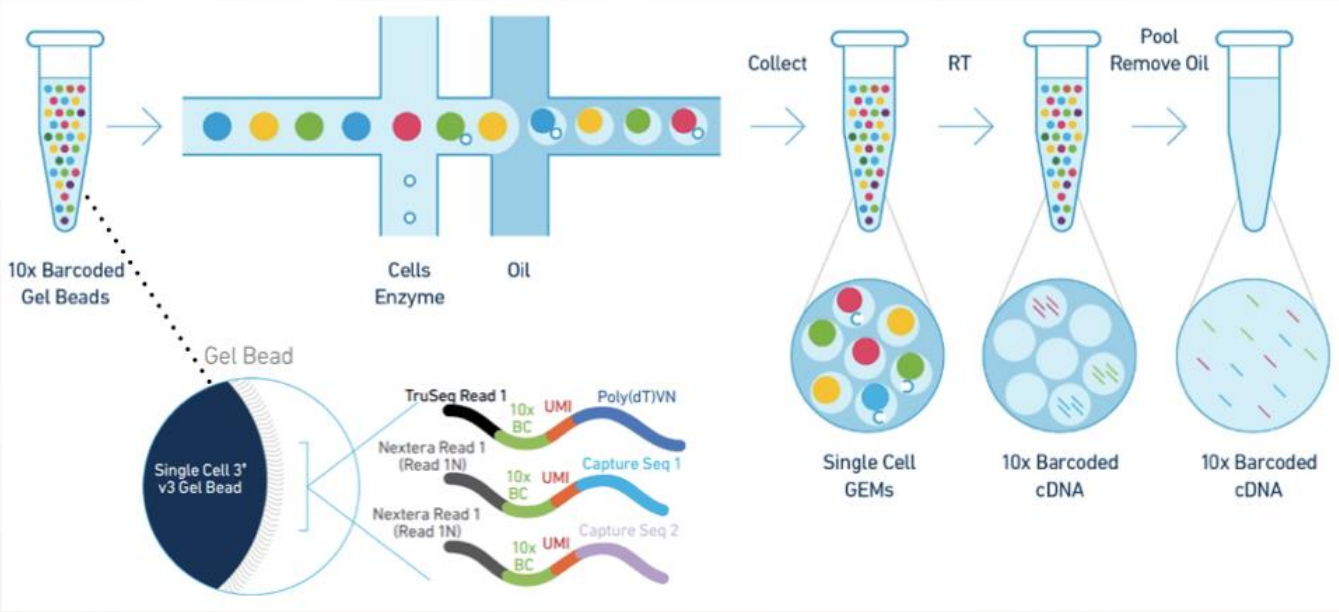


Authors
Evan Z. Macosko, Anindita Basu, ..., Aviv Regev, Steven A. McCarroll

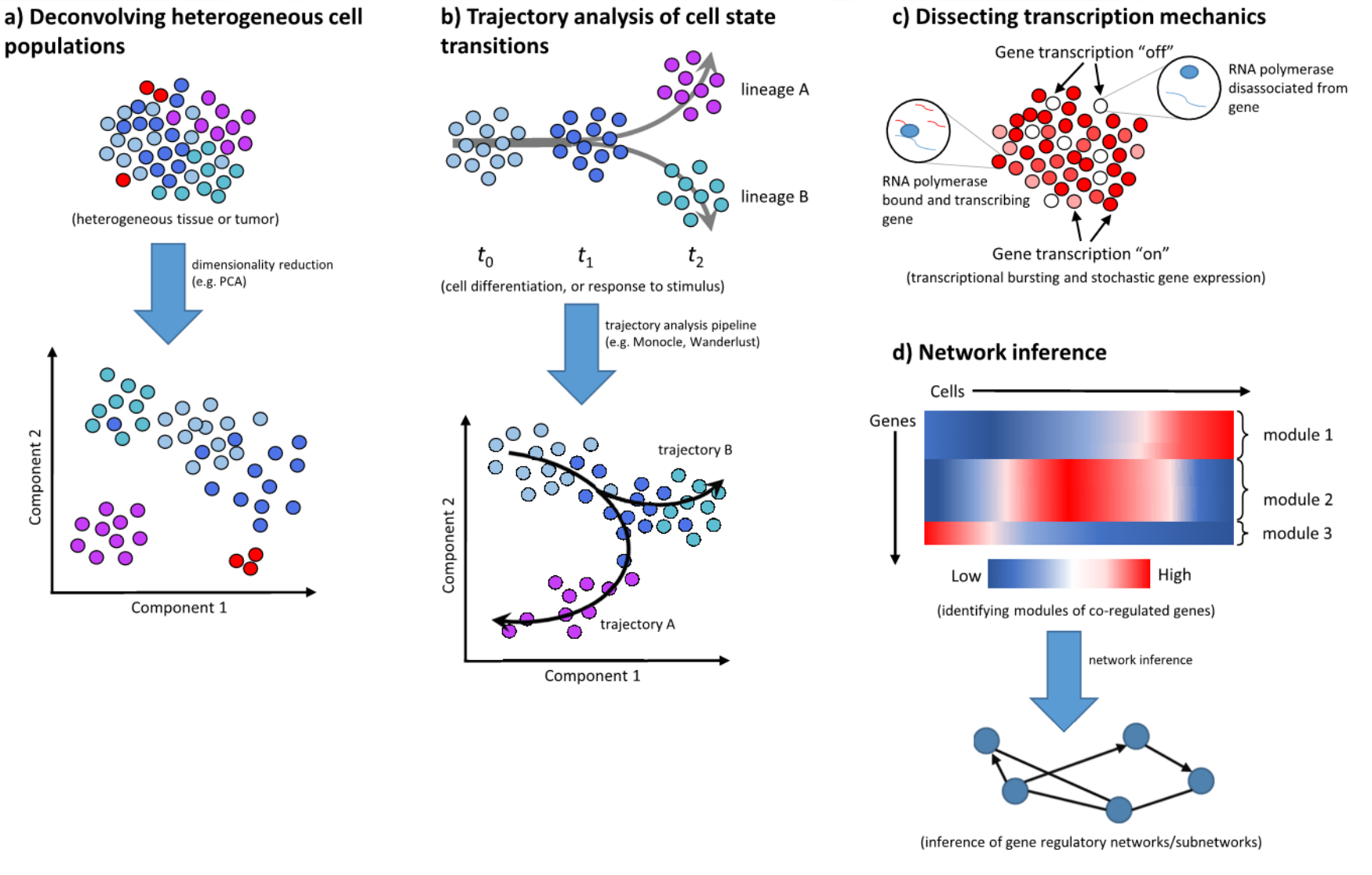
Correspondence
emacosko@genetics.med.harvard.edu (E.Z.M.), mccarroll@genetics.med.harvard.edu (S.A.M.)

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.

10X

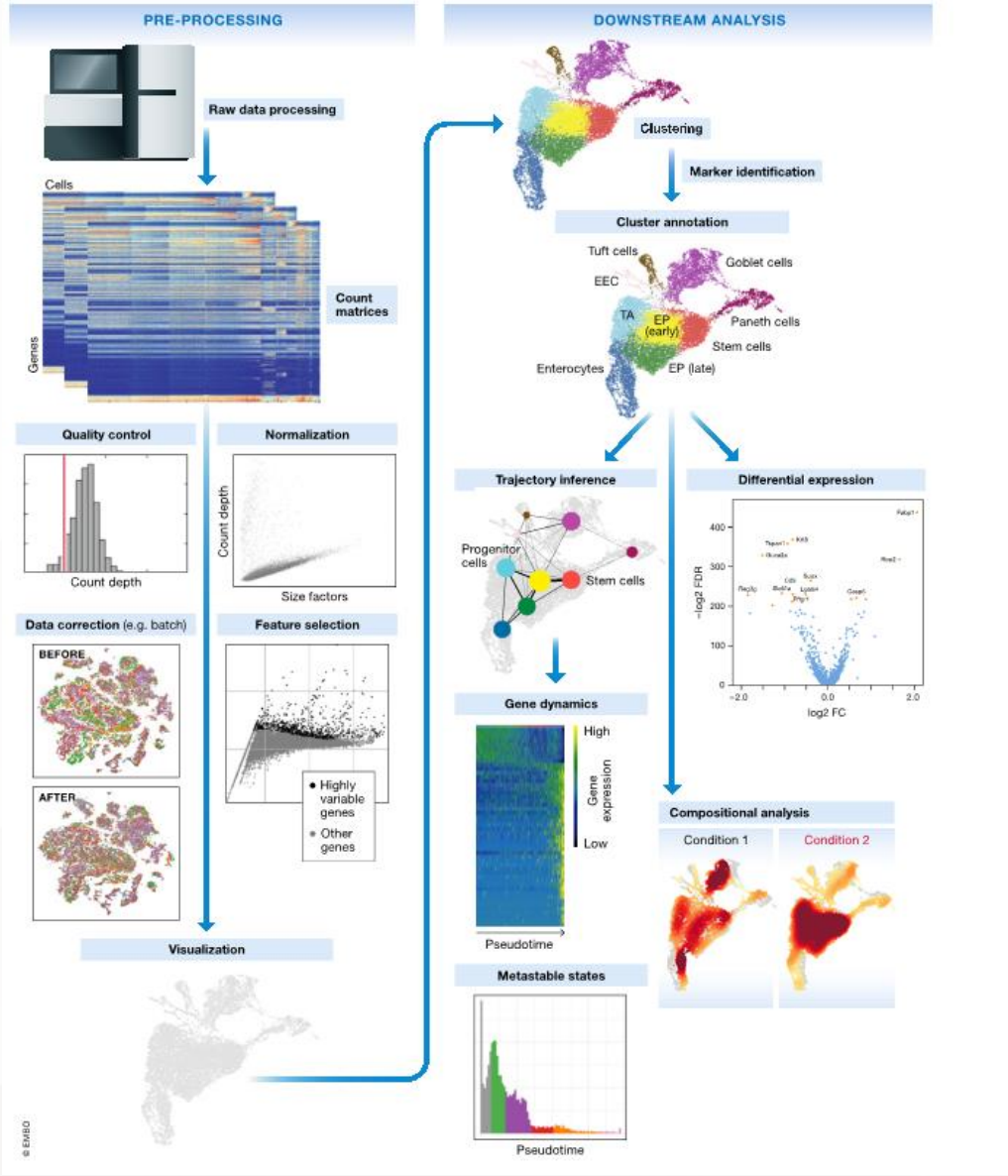


Common applications of scRNA-seq



Single-cell transcriptome sequencing: recent advances and remaining challenges

Schematic of a typical single-cell RNA-seq analysis workflow.



Current best practices in single-cell RNA-seq analysis: a tutorial

上海科技大学
ShanghaiTech University

Outline

- cellranger: 上游分析
- Seurat： 下游分析
- 资源分享+答疑

以cellranger-6.1.2版本为例：

安装cellranger: 从10X官网安装 （https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest)

首先进入下载存放的路径，然后解压

```
[hpc_train@HPC-login 0-software]$ cd /hpc/data/home/hpc_train/sist_2021/songmf/scRNA/0-software
[hpc_train@HPC-login 0-software]$ tar xzvf cellranger-6.1.2
```

将可执行文件，加入环境变量

```
[hpc_train@HPC-login cellranger-6.1.2]$ vim ~/.bashrc
```

```
export PATH=/hpc/data/home/hpc_train/sist_2021/songmf/scRNA/0-software/cellranger-6.1.2:$PATH
```

```
[hpc_train@HPC-login cellranger-6.1.2]$ source ~/.bashrc
```

10X原始数据组成

原始数据：

```
[hpc_train@HPC-login cellranger_tiny_fastq]$ cd /hpc/data/home/hpc_train/sist_2021/songmf/scRNA/0-software/cellranger-6.1.2/external/cellranger_tiny_fastq
[hpc_train@HPC-login cellranger_tiny_fastq]$ ll
total 39388
-rwxr-xr-x 1 hpc_train hpc_lab  3243650 Oct 24 22:47 tinygex_S1_L001_I1_001.fastq.gz
-rwxr-xr-x 1 hpc_train hpc_lab  6655188 Oct 24 22:47 tinygex_S1_L001_R1_001.fastq.gz
-rwxr-xr-x 1 hpc_train hpc_lab 10228498 Oct 24 22:47 tinygex_S1_L001_R2_001.fastq.gz
-rwxr-xr-x 1 hpc_train hpc_lab  3247850 Oct 24 22:47 tinygex_S1_L002_I1_001.fastq.gz
-rwxr-xr-x 1 hpc_train hpc_lab  6654446 Oct 24 22:47 tinygex_S1_L002_R1_001.fastq.gz
-rwxr-xr-x 1 hpc_train hpc_lab 10296017 Oct 24 22:47 tinygex_S1_L002_R2_001.fastq.gz
```

R1: cellbarcode + UMI

R2: 测序片段

I1：sample index序列

参考基因组：

```
[hpc_train@HPC-login cellranger_tiny_ref]$ cd /hpc/data/home/hpc_train/sist_2021/songmf/scRNA/0-software/cellranger-6.1.2/external/cellranger_tiny_ref
[hpc_train@HPC-login cellranger_tiny_ref]$ ll
total 16
drwxr-xr-x 2 hpc_train hpc_lab 4096 Oct 24 22:47 fasta
drwxr-xr-x 2 hpc_train hpc_lab 4096 Oct 24 22:47 genes
-rwxr-xr-x 1 hpc_train hpc_lab  391 Oct 24 22:47 reference.json
drwxr-xr-x 2 hpc_train hpc_lab 4096 Oct 24 22:47 star
```

备注：请根据自己课题，使用恰当的参考基因组！

cellranger count基础命令

请大家先将1-RunCellrangerCount.pbs，拷贝至自己目录下

```
[hpc_train@HPC-login 1-countMatrix]$ cd /hpc/data/home/hpc_train/sist_2021/songmf/scRNA/1-countMatrix
[hpc_train@HPC-login 1-countMatrix]$ cp 1-RunCellrangerCount.pbs ./
```

```
[hpc_train@HPC-login 1-countMatrix]$ cd /hpc/data/home/hpc_train/sist_2021/songmf/scRNA/1-countMatrix
[hpc_train@HPC-login 1-countMatrix]$ cat 1-RunCellrangerCount.pbs
#PBS -N PBS_tiny
#PBS -l nodes=1:ppn=1
#PBS -l walltime=5:00:00
#PBS -S /bin/bash
#PBS -q pub_jx

echo $(date +%R)

cd $PBS_O_WORKDIR
REF="/hpc/data/home/hpc_train/sist_2021/songmf/scRNA/0-software/cellranger-6.1.2/external/cellranger_tiny_ref"
FASTQ="/hpc/data/home/hpc_train/sist_2021/songmf/scRNA/0-software/cellranger-6.1.2/external/cellranger_tiny_fastq"
cellranger count --id=tinygex \
--fastqs=${FASTQ} \
--transcriptome=$REF \
--sample=tinygex \
--localcores=1

echo $(date +%R)
```

参数详解：

--id : output name

--fastqs : fastq files path

--transcriptome : reference path

--sample : the prefix of fastq

--localcores : max cores

cellranger count –help ： 查看更多参数使用

cellranger count结果输出

```
[hpc_train@HPC-login tinygex]$ cd /hpc/data/home/hpc_train/sist_2021/songmf/scRNA/1-countMatrix/tinygex/
[hpc_train@HPC-login tinygex]$ ll outs/filtered_feature_bc_matrix/
total 56
-rwxr-xrwx 1 hpc_train hpc_lab  6439 Nov 19 18:39 barcodes.tsv.gz
-rwxr-xrwx 1 hpc_train hpc_lab  2529 Nov 19 18:39 features.tsv.gz
-rwxr-xrwx 1 hpc_train hpc_lab 43457 Nov 19 18:39 matrix.mtx.gz
```

结果输出：

barcodes : cell barcodes
features : gene name
matrix : UMI


Q & A

Seurat下游分析

Seurat的安装情

请将1-run-seurat.R和0-run-seurat.pbs拷贝至自己目录

```
[hpc_train@HPC-login 2-pbmc]$ cd /hpc/data/home/hpc_train/sist_2021/songmf/scRNA/2-pbmc
[hpc_train@HPC-login 2-pbmc]$ cp 0-run-seurat.pbs 1-run-seurat.R ./
```

```
[hpc_train@HPC-login 2-pbmc]$ cd /hpc/data/home/hpc_train/sist_2021/songmf/scRNA/2-pbmc
[hpc_train@HPC-login 2-pbmc]$ ll -t
total 30184
-rw------- 1 hpc_train hpc_lab     1367 Nov 21 13:33 PBS_pbmc.o3058867
-rw------- 1 hpc_train hpc_lab    10880 Nov 21 13:33 PBS_pbmc.e3058867
-rw-r--r-- 1 hpc_train hpc_lab  1833563 Nov 21 13:33 1-seurat-output.pdf
-rw-r--r-- 1 hpc_train hpc_lab     1890 Nov 21 13:22 1-run-seurat.R
-rw-r--r-- 1 hpc_train hpc_lab      198 Nov 21 13:21 0-run-seurat.pbs
-rw-r--r-- 1 hpc_train hpc_lab    45900 May 27  2016 barcodes.tsv
-rw-r--r-- 1 hpc_train hpc_lab   816952 May 27  2016 genes.tsv
-rw-r--r-- 1 hpc_train hpc_lab 28179564 May 27  2016 matrix.mtx
```

Seurat安装

此处以conda包的管理安装Seurat

创建环境

```
[hpc_train@HPC-login 2-pbmc]$ conda create -n scRNA
```

激活环境

```
[hpc_train@HPC-login 2-pbmc]$ source activate scRNA
```

安装

```
(scRNA) [hpc_train@HPC-login 2-pbmc]$ conda install -c conda-forge r-seurat
```

1-run-seurat.R

Part I

```
[hpc_train@HPC-login 2-pbmc]$ cat 1-run-seurat.R
library(Seurat)
library(dplyr)
library(patchwork)

rm(list=ls())

pdf('1-seurat-output.pdf')
# Load the PBMC dataset
pbmc.data <- Read10X(data.dir = "/hpc/data/home/hpc_train/sist_2021/songmf/scRNA/2-pbmc/")
str(pbmc.data)

# Initialize the Seurat object with the raw (non-normalized data).
pbmc <- CreateSeuratObject(counts = pbmc.data, project = "pbmc3k")

# The [[ operator can add columns to object metadata. This is a great place to stash QC stats
pbmc[["percent.mt"]] <- PercentageFeatureSet(pbmc, pattern = "^MT-")

# Visualize QC metrics as a violin plot
VlnPlot(pbmc, features = c("nFeature_RNA", "nCount_RNA", "percent.mt"), ncol = 3)

#  check correlation
plot1 <- FeatureScatter(pbmc, feature1 = "nCount_RNA", feature2 = "percent.mt")
plot2 <- FeatureScatter(pbmc, feature1 = "nCount_RNA", feature2 = "nFeature_RNA")
plot1 + plot2

# remove low quality cells
pbmc <- subset(pbmc, subset = nFeature_RNA > 200 & nFeature_RNA < 2500 & percent.mt < 5)

# Normaliztion
pbmc <- NormalizeData(pbmc, normalization.method = "LogNormalize", scale.factor = 10000)
```

Part II

```
# find HGV
pbmc <- FindVariableFeatures(pbmc, selection.method = "vst", nfeatures = 2000)

# scaleing Data
all.genes <- rownames(pbmc)
pbmc <- ScaleData(pbmc, features = all.genes)
#pbmc <- ScaleData(pbmc) only for 2000 HVGs

# linear dimensional reduction
pbmc <- RunPCA(pbmc, features = VariableFeatures(object = pbmc))
# how to select sig PC
ElbowPlot(pbmc)

# embeding in two demension
pbmc <- RunUMAP(pbmc, dims = 1:10)
DimPlot(pbmc)

# cluster Cells
pbmc <- FindNeighbors(pbmc, dims = 1:10)
pbmc <- FindClusters(pbmc, resolution = 0.5)

DimPlot(pbmc,label = T)

# find clutter marker genes
pbmc.markers <- FindAllMarkers(pbmc, only.pos = TRUE, min.pct = 0.25, logfc.threshold = 0.25)
pbmc.markers %>% filter(cluster == 3) %>% head()
FeaturePlot(pbmc,'CD79A')
FeaturePlot(pbmc,c('CD3D','CD4','CD8A','CD14'))

dev.off()
```

可变步骤：

- **Remove low quality cells (根据自己课题，去除低质量细胞)**
- **embeding in two dimension （请根据elbow图，去先择合适的PC数目，可选在拐点处)**
- **cluster Cells： 可根据调节resolution的大小，决定cluster的数目。Resolution越大，cluster越多。**

0-run-seurat.pbs

```
[hpc_train@HPC-login 2-pbmc]$ cat 0-run-seurat.pbs
#PBS -N PBS_pbmc
#PBS -l nodes=1:ppn=1
#PBS -l walltime=5:00:00
#PBS -S /bin/bash
#PBS -q pub_jx

echo $(date +%R)

# my seurat package under this conda envs, please according to your seurat installation to change
source activate scRNA

cd $PBS_O_WORKDIR
Rscript 1-run-seurat.R

echo $(date +%R)
```

结果输出

保存的绘图：　　1-seurat-output.pdf
作业error信息：　PBS_pbmc.e3058867
作业output信息：PBS_pbmc.o305886


Q & A

资源分享

- Broad研究所workshop:https://broadinstitute.github.io/2019_scWorkshop/
- Seurat教程：　https://satijalab.org/seurat/articles/pbmc3k_tutorial.html
- 在线单细胞数据：https://singlecell.broadinstitute.org/single_cell
- 单细胞类型marker: http://biocc.hrbmu.edu.cn/CellMarker/


Thanks!