题目：scRNA-seq探究猪核移植胚胎二细胞期发育情况

立项人：吉林大学 朱明远

项目意义：

体细胞核移植技术（SCNT）是将细胞核注射到去核卵母细胞中形成重组的克隆胚胎，将终末分化的体细胞重编程为具有全能性的胚胎细胞，在发育生物学的基础研究以及模式动物的构建方面都具有很大的潜力。虽然体细胞核移植胚胎相比较于诱导多能干细胞，端粒长度、DNA甲基化修饰、主蛋白修饰以及基因表达水平上都更加完美，是全能性细胞，但核移植仍存在许多问题。体细胞核移植的效率很低，大量的克隆胚胎在2-细胞期或4-细胞期被阻滞而无法发育为囊胚，只有少量的胚胎移植到子宫中可以继续发育并出生。此外，许多克隆动物存在发育异常，包括肥胖、胎盘过大以及过早死亡等。种种问题表明，核移植胚胎的重编程可能是不正常或者不完全的。

要研究核移植失败的原因、提高目前克隆方法的效率，必须对供体基因组早期胚胎发育阶段重编程的分子机制进行更广泛深入的研究。研究表明，核移植胚胎在DNA甲基化去除、胚胎基因激活以及组蛋白修饰等方面都存在异常，例如在核移植胚胎的二细胞期仍有体细胞特异性基因的表达，囊胚期仍有许多像Oct4和Sox2等全能性相关基因没有被激活-这些因素的累积作用可能对克隆胚胎或胎儿妊娠期发育产生毒害，严重影响植入后胚胎的发育。在后续对核移植技术的改进研究中，研究者发现虽然药物的使用可以提高克隆胚胎到囊胚的发育率，，但由于其毒性以及对克隆胚胎后续发育的不确定性，实际应用方面受到了很大的限制。因此，如何有效的提高克隆效率，是模式动物建立工作中亟需解决的问题。迄今为止，还不清楚克隆胚胎遗传重编程的精确机制，对克隆胚胎发育分子机制的研究，将会促进对植入前胚胎及重编程研究的发展。

之前的常规的转录组分析方法研究胚胎发育，基于成百上千个核移植胚胎的混合样本进行整体分析，并不能区分发育正常与被阻滞的胚胎，因此可能会掩盖核移植胚胎发育过程中的重要信息，且无法捕获胚胎植入前发育前几天的详细情况。为了精确研究核移植胚胎发育异常的原因，我们采用了单细胞转录组测序并构建了相关分析流程，在原有的计算分析方法上调整、开发新的方法，来适应本次试验数据的特异性。我们希望通过这项研究对比正常发育和核移植胚胎在二细胞期基因表达的不同，更好的了解体细胞核重编程机制以及主要阻滞因素，为进一步校正关键基因的表达来提高核移植效率提供理论依据。

内容：

本次试验将构建好的猪卵丘细胞核移植胚胎以及人工授精胚胎，在体外发育到2-cell阶段，收集样品，进行单细胞转录组测序。设计Linux服务器流程处理数据，运用聚类、维恩图以及热图等方法，对单细胞数据进行可视化分析，揭示胚胎2-cell时期转录组差异。

1.1胚胎及卵细胞的获取

1.1 体外受精胚胎与核移植的收集与培养

8到9月龄的猪于发情13、14或15天肌肉注射氯前列腺烯醇100ug，8小时后再注射同等剂量的氯前列烯醇。在第一次注射氯前列腺烯醇后24小时，肌肉注射PMSG 750-1000IU。72小时后，肌肉注射500-1000IUhCG。注射hCG后38到50小时排卵。公猪精液采集后，进行活力检查，人工授精。

采用正冲法由伞部注射冲卵液，并在子宫角上端插管收集冲卵液。把收集到的冲卵液在集卵杯中静止10分钟，待胚胎沉淀后移去上层液并把含有胚胎的液体倒入带有透明角质酸酶的平皿中消化后，显微镜下检查。用口吸管将卵细胞转入培养滴清洗，转入到新的培养基中放入37℃，5%二氧化碳培养箱培养。

核移植胚胎来自于实验室保存。

1.2核移植胚胎的收集

（1）、提前配置好G1培养基，压孔器后放入37℃，5%二氧化碳培养箱中备用

（2）、将核移植与人工授精的2-细胞期的胚胎取出，蛋白酶消化去掉透明带

（3）、再放到HCZB洗三次以上，去除残留的蛋白酶。

（4）、将胚胎逐个放入培养基中静置一会后，用口吸管反复吹打胚胎，使2-细胞的两个卵裂球分开，放在一起。

（5）、每个胚胎收集一个卵裂球用于转录组测序，剩下的一个卵裂球放入扎好小孔的G1培养基中，标记好收集的卵裂球与培养胚胎之间的对应关系。

（6）、定时观察培养卵裂球的发育情况，并记录发育停止时间

2单细胞转录组测序

（1）试验前订购细胞裂解液，本次试验采用Smart-seq2配套的的单细胞全转录组裂解液，保存在-20度条件下。

（2）将收集到的单细胞用0.5%的BSA-PBS溶液清洗三遍，去除多余的杂质污染

（3）用口吸管吸取单个卵裂球，吹入带有裂解液的Eppendorf管中，随细胞带入液体的体积不超过1ul，两组重复数各5个。高速震荡30s。

（4）4℃，1000g离心一分钟，将离心后的样品迅速放到冰上。

（5）PCR仪70℃孵育90s，震荡10s后，充分裂解，液氮运送至公司测序。

3 核移植胚胎的单卵裂球转录组测序

与正常受精的胚胎相比，核移植胚胎中会有很大一部分在2-细胞期被阻滞，为了检测不同发育命运的胚胎在基因表达的不同，我们首先将2-细胞的胚胎分裂成单个卵裂球并取其中的一个进行单细胞转录组测序，剩下的一个继续培养，并观察最终的发育情况-在继续培养中停止分裂或者继续发育形成囊胚。

测序得到的数据通过FTP协议传输至服务器，数据处理在高性能服务器上进行，绘图部分通过网页访问服务器rstudio-server调用8787端口调试图形代码。

3.1 fastqc进行质量控制

获得单细胞RNA-seq数据后，首先使用FastQC检查测序数据的质量，对测序数据进行质量评估（Quality Control），运行结果表示为zip文件以及 html文件，浏览查看报告并制定质控措施。

3.2测序数据的修剪、拆解以及对齐映射

通过Trim Galore（0.4.5版本）对数据进行修剪、去接头以及去除末端修复效应。得到我们的clean dada，数据处理后返回用FastQC检查质量，并作出测序质量的分布图。

测序片段类似“接头—条形码—插入片段—接头”形式，因此我们使用现有的pipeline将条形码进行去除，或者用perl脚本进行处理。Pipeline脚本代码如下，储存在GitHub：

https://github.com/zhumingyuan123/scRNASeqPipeline-master

首先创建susScr11的参考基因组索引，把事先处理好了数据使用subjunc的默认参数对齐映射（alignment）到susScr11参考基因组上，匹配找到最长的匹配序列。使用featureCounts将每个样品的基因表达定量为FPKM（片段每千万碱基对外显子模型的百万碱基读数）以消除测序深度和转录物长度的影响。

3.5表达矩阵的建立

表达矩阵可以称之为scrna-seq数据处理的核心。通常来说，每一行代表一个基因而每一列代表一个细胞，每一个单元格代表着某一的细胞特定的基因的表达水平。数据处理前我们用FastQC 或者 Kraken来进行质量控制同时使用IGV或seqmonk可视化。最后加载R 包SingleCellExperiment以及scater进行操作，获得表达矩阵。基因表达被量化后，数据被整理成为表达矩阵，其中每行对应于基因（或转录本），并且每列对应于单个细胞。

3.6表达矩阵的过滤

表达矩阵生成后，检查该矩阵质量，以去除在对齐映射后质量控制步骤中未检测到的低质量细胞。在这个阶段未能去除的低质量细胞会增加技术噪音，这有可能干扰我们下游寻找差异信号。

由于单细胞转录组数据的质量控制阈值根据实验设计而定，目前没有处理scRNASeq的统一标准方法。因此为了进行质量控制，我们将筛选相对于其他数据集而言异常的细胞，而不是与某一特定的标准进行比较。

我们使用的scater是在一组QC指标上进行PCA（组成分分析），然后使用自动的异常值检测程序来识别可能存在问题的单元格。

Scater在默认参数下，PCA异常值检测基于以下指标：pct\_counts\_top\_100\_features、total\_features、pct\_counts\_feature\_controls、n\_detected\_feature\_controls、log10\_counts\_endogenous\_features、log10\_counts\_feature\_controls。

scater首先创建一个矩阵， PCA创建了通过其质量度量排序的二维坐标，然后使用mvoutlier包检测异常值。

同理，在进行对数转换后，t-SNE法也可以降维进行质量控制。

最后我们这里使用维恩图比对交集、补集，根据经验进行取舍。完成后使用plotQC函数绘图，以保证转录组的良好覆盖。

3.7 候选基因的筛选

目前已有的寻找单细胞转录组测序数据中的重要基因（feature selection）的方法都不够好，比如 scLVM 主要是根据先验基因集，比如cell-cycle or apoptosis来区分细胞。与此相反，基于 highly variable genes (HVG) 的方法挑选到的变化量大的那些基因很可能是技术带来的误差。而且低表达量基因的变动往往大于高表达量基因，而且所谓的表达变化大也并没有很好的生物学解释。像PCA或者t-SNE这样的降维方法也可以用来挑选重要基因，但它们也受制于系统误差或者批次误差。数据的特异性也会有所影响，dropout是scRNASeq数据的一大特点，就是很多基因在某些细胞根本就不表达，但是在另外的细胞却高表达。综上所述，以往的方法不完全适用于单细胞转录组的数据处理，使用时要加以权衡。

我们选取R包 M3Drop对文库大小进行质量控制和归一化，从而去除检测到的基因较少的细胞和去除未检测到的基因，并将原始读数转换为CPM。然后使用函数rowMeans和rowVars绘制此数据集中所有基因的平均表达和方差之间的关系。

（1）Brennecke\_getVariableGenes法筛选

Brennecke\_getVariableGenes包含在M3Drop中，来校正方差和平均表达量之间的关系。使用前需对文库大小进行归一化，然后计算平均值和平方变异系数（变异除以平均平均表达式）

（2） 筛选高Dropout率基因

Dropout是单细胞数据的主要特征之一，多来自于转录失败。M3Drop的M3DropFeatureSelection函数可以模拟该转录过程，对不符合该模型的数据进行筛选。

（3）scfind查看感兴趣的基因

scfind是一种允许使用基因列表搜索单细胞RNA-Seq的工具，例如搜索表达特定基因的细胞和细胞类型集合。因此我们可以基于已有的研究，手动查看已知的影响胚胎发育的基因。

3.8设计Seurat 流程对数据处理流程进行封装

7、聚类分析

8、功能注释

为了快速有效的进一步识别基因，对候选基因进行功能注释，对差异基因的功能进行注释，就可以发现表达量差异和生物学意义之间的关系。

(1)使用R包 clusterProfiler（http://bioconductor.org/packages/release/bioc/html/clusterProfiler.html）进行注释，注释的数据库为GO/KEGG/REACTOME/MSIGDB。

或者使用网页工具[Variant Effect Predictor](http://asia.ensembl.org/Homo_sapiens/Tools/VEP?db=core)（http://asia.ensembl.org/Homo\_sapiens/Tools/VEP?db=core）+g：profiler（https://biit.cs.ut.ee/gprofiler/index.cgi）或Metascape（<http://metascape.org/gp/index.html#/main/step1>进行注释。

(2)转录组功能的变化根据对该时期特异性高表达的基因进行GO分析，确定P-value最显著的三个GO功能

方案进度：2018.7-2018.8，进行二细胞期样品的采集，送测序；

* + - 1. 样品的质控、对齐、表达矩阵的获取以及质控；

2018.10-2018.12,去干扰、差异表达、组成分分析、通路富集；

2019.1-2019.3；表观遗传相关基因、细胞周期相关基因以及新的转录本基因差异比对；

* 1. 结题报告的撰写。

预期成果（包含项目创新点及特色）

本次试验采用是单细胞转录组全长测序方案，因此会得到远高于芯片技术所检测到的基因数目。通过对核移植胚胎 2-cell发育阻滞的问题进行的转录组分析，我们比对核移植胚胎与体内正常胚胎之间的转录组数据之间的差异，希望能够探究出潜在影响发育的差异基因，为进一步研究2-cell胚胎的发育机制提供线索。

1. 通过对体内正常细胞、核移植细胞的深入研究，按照转录组不同RPKM值为临界点，筛选出两组中高表达的差异基因，作为候选基因，通过维恩图、热图、GO和KEGG注释等分析方法，得到显著差异表达基因
2. 两组胚胎发育的差异基因，以一定的差异系数筛选出差异基因作为候选分析基因
3. 分析表观遗传相关基因、细胞周期相关基因以及新的转录本基因，比较体内胚胎与核移植胚胎在这些基因表达水平上的差异

项目的特色以及创新点：

### 本次试验采取了以scater包为核心转录组测序流程进行数据处理，得到表达矩阵，进行差异分析。传统的 Tophat+Cufflinks流程、RSEM, express,Sailfish与本方案相比，传统方法多适用于bulk RNA-seq，这种数据是成千上万的细胞的混合数据，没有完全考虑到seRNA-seq数据高dropouts等特点

# References

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. “Basic Local Alignment Search Tool.” Journal of Molecular Biology 215 (3). Elsevier BV: 403–10. doi:10.1016/s0022-2836(05)80360-2.

Anders, Simon, and Wolfgang Huber. 2010. “Differential Expression Analysis for Sequence Count Data.” Genome Biol 11 (10). Springer Nature: R106. doi:10.1186/gb-2010-11-10-r106.

Archer, Nathan, Mark D. Walsh, Vahid Shahrezaei, and Daniel Hebenstreit. 2016. “Modeling Enzyme Processivity Reveals That RNA-Seq Libraries Are Biased in Characteristic and Correctable Ways.” Cell Systems 3 (5). Elsevier BV: 467–479.e12. doi:10.1016/j.cels.2016.10.012.

Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. “Fast Unfolding of Communities in Large Networks.” J. Stat. Mech. 2008 (10). IOP Publishing: P10008. doi:10.1088/1742-5468/2008/10/p10008.

Bray, Nicolas L, Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. “Near-Optimal Probabilistic Rna-Seq Quantification.” Nat Biotechnol 34 (5): 525–27. doi:10.1038/nbt.3519.

Bullard, James H, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit. 2010. “Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiments.” BMC Bioinformatics 11 (1). Springer Nature: 94. doi:10.1186/1471-2105-11-94.

Buttner, Maren, Zhichao Miao, Alexander Wolf, Sarah A Teichmann, and Fabian J Theis. 2017. “Assessment of Batch-Correction Methods for scRNA-seq Data with a New Test Metric.” bioRxiv, October, 200345.

Cannoodt, Robrecht, Wouter Saelens, and Yvan Saeys. 2016. “Computational Methods for Trajectory Inference from Single-Cell Transcriptomics.” Eur. J. Immunol. 46 (11). Wiley-Blackwell: 2496–2506. doi:10.1002/eji.201646347.

Deng, Q., D. Ramskold, B. Reinius, and R. Sandberg. 2014. “Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells.” Science 343 (6167). American Association for the Advancement of Science (AAAS): 193–96. doi:10.1126/science.1245316.

Dijk, David van, Juozas Nainys, Roshan Sharma, Pooja Kathail, Ambrose J Carr, Kevin R Moon, Linas Mazutis, Guy Wolf, Smita Krishnaswamy, and Dana Pe’er. 2017. “MAGIC: A Diffusion-Based Imputation Method Reveals Gene-Gene Interactions in Single-Cell RNA-sequencing Data.” bioRxiv, February, 111591.

Gierahn, Todd M, Marc H Wadsworth 2nd, Travis K Hughes, Bryan D Bryson, Andrew Butler, Rahul Satija, Sarah Fortune, J Christopher Love, and Alex K Shalek. 2017. “Seq-Well: Portable, Low-Cost RNA Sequencing of Single Cells at High Throughput.” Nat. Methods 14 (4): 395–98.

Guo, Minzhe, Hui Wang, S. Steven Potter, Jeffrey A. Whitsett, and Yan Xu. 2015. “SINCERA: A Pipeline for Single-Cell RNA-Seq Profiling Analysis.” PLoS Comput Biol 11 (11). Public Library of Science (PLoS): e1004575. doi:10.1371/journal.pcbi.1004575.

附录：

基于以上对软件以及方法的选择，我们制定了一个基于seurat包的流程。Seurat最初是作为scRNA-seq数据的集群处理工具，目前它发展成为一个广泛使用的R软件包，可以对scRNA-seq数据进行质量控制，分析和探索。我们主要在各种函数调用中使用默认值，具体的调用软件就是我们之前所探讨的部分。

流程实验脚本（未进行管道化处理）

### 6.1 导入数据

SCE <- readRDS("文件名.rds")

##################################

### 6.2 创建Seurat对象

library(SingleCellExperiment)

library(Seurat)

library(mclust)

library(dplyr) ### 载入R包

seuset <- CreateSeuratObject(

raw.data = counts(SCE),

min.cells = 3,

min.genes = 200

)

##########################################

### 6.3 表达数据的质量控制

####

VlnPlot(

object = seuset,

features.plot = c("nGene", "nUMI"),

nCol = 2

)

### 将基因和分子计数可视化并绘制它们之间的关系

GenePlot(

object = seuset,

gene1 = "nUMI",

gene2 = "nGene"

)

seuset <- FilterCells(

object = seuset,

subset.names = c("nUMI"),

high.thresholds = c(2e7)

)

### 排除具有明确异常值读取次数的单细胞

##################################

### 6.4标准化

### 全局缩放归一化方法LogNormalize，通过总表达式对每个细胞的基因表达测量结果进行归一化，将其乘以比例因子（默认为10,000），并对结果进行对数转换

seuset <- NormalizeData(

object = seuset,

normalization.method = "LogNormalize",

scale.factor = 10000

)

###########################

### 6.5 筛选出差异基因

###### Seurat计算高度可变的基因并将其用于下游分析。#####FindVariableGenes计算每个基因的平均表达和离散度，这有助于控制变异性和平均表达之间的关系

seuset <- FindVariableGenes(

object = seuset,

mean.function = ExpMean,

dispersion.function = LogVMR,

x.low.cutoff = 0.0125,

x.high.cutoff = 3,

y.cutoff = 0.5

)

#################

### 6.6 处理其他干扰因素

#### Seurat可以通过批次，细胞比对率（Drop-seq工具提供的Drop-seq数据），检测到的分子数量，线粒体基因表达和细胞周期来推断出基因表达的细胞 - 细胞变异。

seuset <- ScaleData(

object = seuset,

vars.to.regress = c("nUMI")

)

###########################

### 6.7 线性降维

### 下来我们PCA对缩放数据执行操作。在高变区基因上运行降维可以提高运行速度。

seuset <- RunPCA(

object = seuset,

pc.genes = seuset@var.genes,

do.print = TRUE,

pcs.print = 1:5,

genes.print = 5

)

PrintPCA(object = seuset, pcs.print = 1:5, genes.print = 5, use.full = FALSE)

VizPCA(object = seuset, pcs.use = 1:2)

PCAPlot(object = seuset, dim.1 = 1, dim.2 = 2)

PCHeatmap(

object = seuset,

pc.use = 1:6,

cells.use = 500,

do.balanced = TRUE,

label.columns = FALSE,

use.full = FALSE

)

#########

### 6.8 细胞聚类

seuset <- FindClusters(

object = seuset,

reduction.type = "pca",

dims.use = 1:8,

resolution = 1.0,

print.output = 0,

save.SNN = TRUE

)

###############

### 6.9 标记候选基因

##Seurat可以标记差异表达聚类的基因。默认参数下，它会识别单个群体中与所有其他单元格相比较上调与下调的部分，并对他们进行标记。通过聚类展示样品之间趋同性与差异性

### 寻找标记为2的部分

markers2 <- FindMarkers(seuset, 2)

### 然后可以将标记基因可视化：

VlnPlot(object = seuset,features.plot=rownames(markers2)[1:2])

FeaturePlot(

seuset,

head(rownames(markers2)),

cols.use = c("lightgrey", "blue"),

nCol = 3

)

### FindAllMarkers 自动执行此过程并查找群体中所有的标记：

markers <- FindAllMarkers(

object = seuset,

only.pos = TRUE,

min.pct = 0.25,

thresh.use = 0.25

)

### DoHeatmap为给定的细胞和基因生成表达式热图。在这种情况下，我们绘制每个集群的前10个标记（如果小于20，则绘制所有标记）：

top10 <- markers %>% group\_by(cluster) %>% top\_n(10, avg\_logFC)

DoHeatmap(

object = seuset,

genes.use = top10$gene,

slim.col.label = TRUE,

remove.key = TRUE

)