Jan 15, 2019 Version 5

Bioinformatic pipeline for studying transcriptome and regulome dynamics during neural differentiation V.5

DOI
dx.doi.org/10.17504/protocols.io.w6dfha6

Zhouchun Shang[1], Dongsheng Chen[1], Quanlei Wang[1], Shengpeng Wang[1]

[1]BGI-shenzhen, Shenzhen, China

Wang Shengpeng

DOI: dx.doi.org/10.17504/protocols.io.w6dfha6

Protocol Citation: Zhouchun Shang, Dongsheng Chen, Quanlei Wang, Shengpeng Wang 2019. Bioinformatic pipeline for studying transcriptome and regulome dynamics during neural differentiation. **protocols.io**
https://dx.doi.org/10.17504/protocols.io.w6dfha6

Protocol status: Working
We use this protocol and it's working

Created: January 15, 2019

Last Modified: January 15, 2019

Protocol Integer ID: 19365

Abstract

Here we present a bioinformatic pipeline for dissecting transcriptional regulation of neural differentiation process. Single cell RNA-seq and bulk ATAC-seq with two biological replicates was applied to the indicated cell stages including human induced pluripotent stem cells (hiPSCs), embryoid body (EB), early rosettes (Ros-E), late rosettes (Ros-L), neural progenitor cells (NPCs), and the original somatic fibroblasts (Fib) for a deeper understanding of the regulatory mechanisms driving the differentiation of the neural lineage. This pipeline could be applied to study transcriptome and regulome dynamics of other lineages.

Raw data description

1 Single cell RNA-seq: 50bp single-end sequencing was performed using the BGISEQ-500 platform
Bulk RNA-seq: 50bp single-end sequencing was performed using the BGISEQ-500 platform
Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) was sequenced on BGISEQ-500 platform

Assess read qualities of read files (fastq files) using FASTQC

2 FASTQC

Pre-processing of raw data

3 The original FASTQ data of the 527 samples were aligned to the rRNA database (downloaded from NCBI) to remove rRNAs and the remaining reads were processed with SOAPnuke (version 1.5.3) to trim adaptors and filter out the low-quality reads. The filtered data were aligned to the reference genome (hg19) using hisat2 (HISAT2 version 2.0.1-beta). Reads were counted using the R package GenomicAlignments (mode='Union', inter.feature=FALSE), and normalized to RPKM with edgeR.

Quality control of single cell RNA-seq

4 Cells were filtered using following parameters: genome mapping rate more than 70%, fraction of reads mapped to mitochondrial genes less than 20%, mRNA mapping rate more than 80%, ERCC ratio less than 10%, and gene number more than 5000. Further, correlation of ERCC among cells was used to evaluate the quality of each cell (threshold=0.9). At last, 445 single cells remained for further analysis in this project.

Identification of differentially expressed genes

5 Differentially expressed genes in iPSCs, EB, Ros-E, Ros-L, and NPCs was determined using SCDE (single cell differential expression analysis) with default parameters except requiring a minimum of 100 genes (parameter min.lib.size = 100 to call scde.error. models function). The Z scores and corrected Z scores (cZ) to adjust for the multiple testing were converted into two-tailed p-values and adjusted to control for FDR using pnorm function in R. The significantly differentially expressed genes were selected based on following criteria: adjusted p-value < 0.01 and fold-change > 2.
###################Detailed script in this step####################

```
da =
read.table('expression_matrix.txt',header=T,row.names=1,check.names=FALSE,sep='\t')
####do Differential Expression
library(methods)
library(scde)
EB=da[,grep('EB',colnames(da))]
all=da[,c(-grep('EB',colnames(da)),-grep('Fib',colnames(da)))]
x = cbind(EB,all)
sg1 ← gsub('(b[0-9]+).(EB|all)*', '\\2', colnames(x))
sg ← factor(gsub('(EB|all).*', '\\1', sg1), levels = c('EB', 'all'))
sg[is.na(sg)] = 'all'
names(sg) ← colnames(x)
table(sg)
cd ← clean.counts(x, min.lib.size=100, min.reads = 1, min.detected = 1)
o.ifm ← scde.error.models(counts = cd, groups = sg, n.cores = 1, threshold.segmentation
= TRUE, save.crossfit.plots = FALSE, save.model.plots = FALSE, verbose = 1)
save(o.ifm,file='EB_VS_all.scde.error.models.RData')
#load('EB_VS_all.scde.error.models.RData')
valid.cells ← o.ifm$corr.a > 0
table(valid.cells)
o.ifm ← o.ifm[valid.cells, ]
o.prior ← scde.expression.prior(models = o.ifm, counts = cd, length.out = 400, show.plot
= FALSE)
group1 ← gsub('(b[0-9]+).(EB|all)*', '\\2', rownames(o.ifm))
group ← factor(gsub('(EB|all).*', '\\1', group1), levels = c('EB', 'all'))
group[is.na(group)] = 'all'
names(group) ← row.names(o.ifm)
ediff ← scde.expression.difference(o.ifm, cd, o.prior, groups = group, n.randomizations =
100, n.cores = 1, verbose = 1)
p.values ← 2*pnorm(abs(ediff$Z),lower.tail=F) # 2-tailed p-value
p.values.adj ← 2*pnorm(abs(ediff$cZ),lower.tail=F) # Adjusted to control for FDR
significant.genes ← which(p.values.adj<0.05)
length(significant.genes)
ord ← order(p.values.adj[significant.genes]) # order by p-value
de ← cbind(ediff[significant.genes,1:3],p.values.adj[significant.genes])[ord,]
colnames(de) ← c('Lower bound','log2 fold change','Upper bound','p-value')
write.table(de, file = 'EB_VS_all.scde.FDR.xls', row.names = TRUE, col.names = TRUE,
sep = '\t', quote = FALSE)
```

Constructing trajectory using variable genes

6 Monocle ordering was conducted for all iPSCs, EB, Ros-E, Ros-L and NPCs cells using the set of variable genes with default parameters except we specified reduction_method

=”DDRTree” in the reduceDimension function. The variable genes were selected using the Seurat R package.

```
##################Detailed script in this step####################
myda = read.table('expression_matrix.txt',sep='\t',header = T,row.names = 1)
exp = subset(exp,select=c(-grep('H9|Fib',colnames(exp))))
library(Seurat)
marrow <- CreateSeuratObject(raw.data = myda)
marrow <- NormalizeData(object = marrow)
marrow <- FindVariableGenes(object = marrow,
mean.function = ExpMean,
dispersion.function = LogVMR,
x.low.cutoff = 0, x.high.cutoff = 100,
y.cutoff = 0.5,
do.plot = TRUE)
dim(x = marrow@var.genes)
myda = myda[as.matrix(marrow@var.genes),]

pheno.data1 <- colnames(myda)
pheno.data2 <- unlist(lapply(pheno.data1, function(x) strsplit(x,'.',fixed = TRUE)[[1]][2]))
pheno.data3 <- unlist(lapply(pheno.data2, function(x) strsplit(x,'_',fixed = TRUE)[[1]][1]))
##remove the quantity id for each samples
pheno.data4 <- unlist(lapply(pheno.data1, function(x) strsplit(x,'_',fixed = TRUE)[[1]][1]))
pheno.data5 <- unlist(lapply(pheno.data4, function(x) strsplit(x,'.',fixed = TRUE)[[1]][1]))
mode(pheno.data5)
pheno.data.df <- data.frame(type=pheno.data3)
rownames(pheno.data.df) <- colnames(myda)
pd <- new('AnnotatedDataFrame', data = pheno.data.df)

feature.fd = rownames(myda)
feature.fd <- unlist(lapply(feature.fd, function(x) strsplit(x, '.')[[1]][1]))
feature.data.fd = data.frame(type=feature.fd)
rownames(feature.data.fd) <- rownames(myda)
fd <- new('AnnotatedDataFrame', data = feature.data.fd)
HSMM <- newCellDataSet(as(as.matrix(myda), 'sparseMatrix'),
phenoData = pd,
featureData = fd,
lowerDetectionLimit=1,
expressionFamily=tobit())
ordering.genes <- rownames(feature.data.fd)
data <- setOrderingFilter(HSMM, ordering.genes)
data <- reduceDimension(data,max_components=2,reduction_method =
c('DDRTree'),norm_method = 'none')
```

```
data <- orderCells(data, num_paths = 2, reverse = FALSE)
save(data,file='monocle_simple_used_391sample.gene.RData')

pdf('monocle_simple_used__main_subtype.gene_subtype.pdf')
plot_cell_trajectory(data, color_by = 'type',show_branch_points = FALSE)
dev.off()
```

Analysis of heterogeneity in each cell stage

7 The heterogeneity of each cell stage was determined using Seurat R package.

ATAC peak calling

8 We aligned ATAC-seq data to hg19 using Bowtie2 and called peaks using MACS2. We established a standard peak set by merging all overlapping peaks. The IDR pipeline was used to identify reproducible peaks between two biological replicates. Only peaks with IDR<0.01 were considered reproducible and retained for downstream analysis. Pearson correlation coefficients of two biological replicates at each stage were calculated. Stage-specific peaks were defined as peaks having no overlap with any peaks in other stages. Novel peaks were defined as peaks non-overlapping with previous stages. In the case of iPSCs, all peaks were annotated as novel peaks.

For reproducible peaks, we applied HOMER to assign putative targets for peaks. For stage-specific peaks, ChIPseeker was used for putative target assignment. In both strategies, the putative target of a certain peak is defined as the gene with TSS closest to the peak summit location.

GO term and KEGG pathway enrichment analysis

9 Lists of genes were analysed using DAVID and the BH method was used for multiple test correction. GO terms with a FDR less than 0.01 or 0.05 were considered as significantly enriched. Target genes of stage-specific ATAC peaks were analysed using the R package, clusterProfiler, in which an adjusted p-value of 0.05 was used to identify significantly enriched GO and KEGG terms associated with each set of peaks.

Regulatory network construction

10 The scRNA-seq profiles among each cell types were compared using SCDE package. TFs significantly differentially expressed, with adjusted p-value threshold of 0.05, among neighboring cell types were submitted to STRING database to infer regulatory networks based on known interaction relationships (supported by data from curated databases, experiments and text-mining). TFs without any interactions with other proteins were

removed from the network. To select key regulators, we used a threshold of 5 and all TFs with number of interactions above the threshold were considered as key regulators.

Construction of cellular communication network

11 The ligand-receptor interaction relationships were downloaded from the database, IUPHAR/BPS Guide to PHARMACOLOGY, and the Database of Ligand-Receptor Partners (DLRP). The average expression level of RPKM of 1 was used as a threshold. Ligands and receptors above the threshold were considered as expressed in the corresponding cluster. The R package Circlize was used to visualize the interactions.

Motif enrichment analysis

12 Motifs enriched in each set of ATAC peaks were identified using findMotifsGenome.pl from HOMER using following parameters: -size -100,100 -len 4,5,6,7,8,9,10,11,12.