RNA-seq data analysis : step-by-step guide

Last updated: 6 June 2019

Data download

✓ `ascp -QT -P33001 -k 1 -i ~/.aspera/connect/etc/asperaweb_id_dsa.openssh era-fasp@link destination`

Quality control (FastQC)

✓ `fastqc file.fastq.gz`

Trimming (Trimmomatic)

✓ `java -jar trimmomatic-0.36.jar PE -trimlog trim.log file_1.fastq.gz file_2.fastq.gz -baseout file_clean.fq.gz ILLUMINACLIP:/Trimmomatic-0.36/adapters/TruSeq3-PE-2.fa:2:30:10 TRAILING:24 MINLEN:85`

Alignment (Hisat2)

✓ `hisat -p 12 --dta -x Human_84/hisat2_index/grch38_tran/genome_tran -1 file_1.fastq.gz -2 file_2.fastq.gz -S file.sam --un-gz file_unaln --summary-file file_sum.txt --met-file file_met.tsv`

SAM to BAM conversion (SAMtools)

✓ `samtools sort -o file.bam file.sam`

Expression estimation by gene counts (FeatureCounts)

✓ `R`

✓ `fc <- featureCounts(bam.list, annot.ext = "file.gtf", isGTFAnnotationFile = T, allowMultiOverlap = F, isPairedEnd=F, nthreads = 10, countMultiMappingReads = F, GTF.featureType = 'gene')`

Differential gene expression (DESeq2)

```
# Read count table and extract counts for Protein coding genes
✓  R
✓  library(DESeq2)
✓  counts = read.table("file.tsv", row.names=c(1), header=T)
✓  protein_coding = as.matrix(read.table("protein_coding"))
✓  counts_pc <- counts[protein_coding,]
# Filtering out genes which donot have 1 cpm atleast in two samples
✓  keep <- rowSums(cpm(counts_pc)>1) >1
✓  counts_pc_filt <- counts_pc[keep, ]
✓  condition = c(rep("T",2), rep("C",2))
✓  coldata <- data.frame(row.names=colnames(counts_pc_filt), condition)
# DESeq2 Analysis
✓  dds = DESeqDataSetFromMatrix(countData=counts_pc_filt, colData=coldata, design=~condition)
✓  dds = DESeq(dds)
✓  res = results(dds)
✓  write.table(res, "DeSeq2_result.tsv", sep="\t", col.names=NA)
✓  res_clean = na.exclude(as.data.frame(res))
✓  upreg = res_clean[(res_clean$log2FoldChange>1 & res_clean$padj<0.01),]
✓  write.table(upreg, "DeSeq2_upreg.tsv", sep="\t", col.names=NA)
✓  downreg = res_clean[(res_clean$log2FoldChange<(-1) & res_clean$padj<0.01),]
✓  write.table(downreg, "DeSeq2_downreg.tsv", sep="\t", col.names=NA)
✓  save (dds, file = DeSeq2_result.rda')
```

Differential gene expression (EdgeR)

```
library(edgeR)
sample_info.edgeR <- factor(c(rep("T", 2), rep("C",2)))
sample_info.edgeR <- relevel(sample_info.edgeR, ref="C")
edgeR.DGElist <- DGEList(counts= counts_pc_filt, group=sample_info.edgeR) # counts_pc_filt from previous section
edgeR.DGElist <- calcNormFactors(edgeR.DGElist, method="TMM")
edgeR.DGElist$samples
design <- model.matrix(~sample_info.edgeR)
edgeR.DGElist <- estimateDisp(edgeR.DGElist, design)
edger_fit <- glmFit(edgeR.DGElist, design)
edger_lrt <- glmLRT(edger_fit)
DGE.results_edgeR <- topTags(edger_lrt, n=Inf, sort.by = "none", adjust.method = "BH")
write.table(DGE.results_edgeR$table, "EdgeR_result.tsv", sep="\t", col.names=NA)
edger.upreg = DGE.results_edgeR$table[(DGE.results_edgeR$table$logFC>1 & DGE.results_edgeR$table$FDR<0.01),]
write.table(edger.upreg, "EdgeR_upreg.tsv", sep="\t", col.names=NA)
edger.downreg = DGE.results_edgeR$table[(DGE.results_edgeR$table$logFC<(-1) & DGE.results_edgeR$table$FDR<0.01),]
write.table(edger.downreg, "EdgeR_downreg.tsv", sep="\t", col.names=NA)
save (edger_lrt, file = 'edgeLRT.rda')
```

Differential gene expression (Limma)

```
rownames(design) <- colnames(edgeR.DGElist) # edgeR.DGElist from previous section
voomTransformed <- voom(edgeR.DGElist, design, plot=FALSE)
voom.fitted <- lmFit(voomTransformed, design = design)
voom.fitted <- eBayes(voom.fitted)
DGE.results_limma <- topTable(voom.fitted, coef="sample_info.edgeRT", number=Inf, adjust.method="BH", sort.by="none")
write.table(DGE.results_limma, "Limma_result.tsv", sep="\t", col.names=NA)
limma.upreg = DGE.results_limma[(DGE.results_limma$logFC>1 & DGE.results_limma$adj.P.Val<0.01),]
write.table(limma.upreg, "Limma_upreg.tsv", sep="\t", col.names=NA)
limma.downreg = DGE.results_limma[(DGE.results_limma$logFC<(-1) & DGE.results_limma$adj.P.Val<0.01),]
write.table(limma.downreg, "Limma_downreg.tsv", sep="\t", col.names=NA)
save (voom.fitted, file = 'VoomFitted.rda')
```