Thema: RNA -Seq

# Name: Dang Quynh Tram Nguyen

### Vorbereitung

* Annotation: wird in <https://www.gencodegenes.org/human/release_39.html> herunterladen, Version nur für superset. Annotation passt mit der Genome File, was in der gleichen Quelle heruntergeladen.

### Splice-aware mapping mit HISAT2

(hisat2 erkennt die Splicing -> besser mappen)

* Hisat2-index aufbauen:

Code: hisat2-build -p 12 <genome.fa> <prefix-Output>

(Speichern alle Output of index in einem Ordner)

* Hisat2 Mapping:

Code: hisat2 -p 8 -x <index> -U <Inputfile> -S <Output-File> --dta -t

--dta : mit StringTie. Effektiveres Alignment um beste Reads zu finden, dabei die möglichsten Intros bestimmt werden.

-t : zeigt die Zeit der Prozess an

-S : Output-File in SAM file

|  |  |  |
| --- | --- | --- |
| **Bezeichnung** | **Insgesamt** | **Eindeutig** **gemappte** **Reads** |
| Control 1 | 65806836 | 50922396 |
| Control 2 | 50453652 | 37680233 |
| E2 1 | 74436420 | 55712934 |
| E2 2 | 52257963 | 36470429 |

* SAM in BAM konvertieren

Code: samtool view -S <Inputfile in SAM> --input-fmt-option nthreads=8 -bo <Output-File in BAM> --output-fmt-option nthreads=8

--(input/output)-fmt-option nthreads=<INT>: Option, um die Input/Output Format mit vielen Threads bearbeitet ((en-)codiert) zu werden

Code: samtools sort <Input in BAM File> --input-fmt-option nthreads=8 -o <Output in BAM File> --output-fmt-option nthreads=8

### Read Counts mit featureCounts

(Die Anzahl der Reads pro Gen)

Code: featureCounts -T 8 -a <Annotation in GTF Format> -o <Output-File in TXT Format> <Inputfile IN BAM>

<Inputfile> : kann ein oder eine List aller Inputfiles in BAM sein

(nutze sortierte BAM-Infutfiles, um Zeit zu sparen)

|  |  |
| --- | --- |
| **Bezeichnungen** | **Read Counts** |
| Control 1 | 11153277 |
| Control 2 | 13913852 |
| E2 1 | 12238113 |
| E2 2 | 13055193 |

* Modifizieren die Output-Files, dabei sie enthalten nur die Gene und ihre Readcounts (falls nur ein vorheriges Inputfile)

Code: cut -f1,7 <Inputfile> | sed ‘1d‘ > <Output-File>

cut -f<INT>: die extrahierende Spalte

sed ‘<INT>d‘: entfernt die <INT>. Zeile

### Scatterplot in R

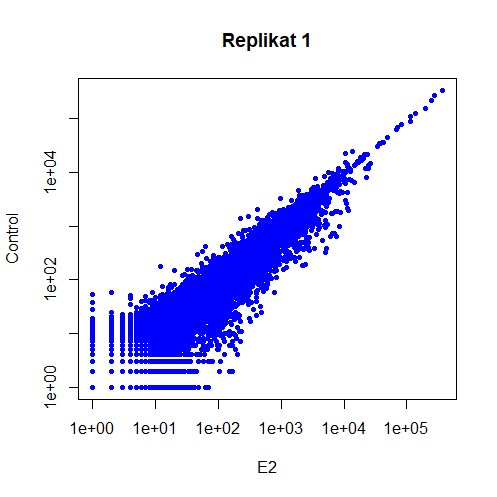
(Verteilung der Reads im Vergleichen zwischen Control und

read.delim(File, header = TRUE, sep = ‘\t‘)

merge(control-File, E2-File)

plot(e2,control, main, xlab, ylab, pch, col,log = „xy“)

Pch : Punkttyp. Pch = 20 (Bullet-punkt)



### Finden differentielle exprimierte Gene mit DESeq2

Referenz:

1. <https://lashlock.github.io/compbio/R_presentation.html>
2. <http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#se>

Code: ressig <- res[which(res$padj < 0.1),]

write.table(as.data.frame(ressig[order(-ressig$log2FoldChange),]),file="hoch- und runterreguliert.txt")

read.table(“hoch- und runterreguliert.txt“)

* Hier gibt es alle Gene mit adjusted p-value unter 0.1. Welche hat positive Fold Change, ist *ein hochreguliertes Gen*, sondern ein *runterreguliertes Gen* mit negativer Fold Change.

out of 41300 with nonzero total read count

adjusted p-value < 0.1

LFC > 0 (up) : 18, 0.044%

LFC < 0 (down) : 1, 0.0024%

outliers [1] : 0, 0%

low counts [2] : 8499, 21%

(mean count < 1)

[1] see 'cooksCutoff' argument of ?results

[2] see 'independentFiltering' argument of ?results

* Würde adjusted p-value statt p-value genutzt, vermeiden die Ergebnisse die FDR (False Discovery Rate) möglichst.

### Vergleichen mit den Ergebnissen von ChIP-Seq

Readcounts und DESeq2 wie obere Schritte.

out of 57711 with nonzero total read count

adjusted p-value < 0.1

LFC > 0 (up) : 1897, 3.3%

LFC < 0 (down) : 1, 0.0017%

outliers [1] : 0, 0%

low counts [2] : 36579, 63%

(mean count < 10)

[1] see 'cooksCutoff' argument of ?results

[2] see 'independentFiltering' argument of ?results