RNA-seq 数据分析报告

摘要

本篇报告分析了来自于人类 Hela 细胞系使用两种 siRNA 对 *hNRNPC* 基因进行敲低的样本与正常的样本之间在基因表达方面的差异，使用的 RNA-seq 高通量测序的方法获取数据，选取了其中 14 号染色体上的数据进行了差异表达分析及主成分与聚类分析，找到了其中差异最为显著的基因。

介绍

本次报告是对 R 语言进行差异基因表达分析课程的练习，下载了 2012 一篇文章的 RNA-seq 数据，选取第 14 条染色体上的基因进行了后续的分析。文章的作者采用 siRNA 敲除的方法敲低的方法降低了 *hnRNPC* 基因的表达，采用了两种不同的 siRNA 进行实验，并使用未经处理的 HeLa 细胞作为对照。

方法

选取了 R 软件 Bioconductor 社区的"RNAseqData.HNRNPC.bam.chr14" 软件包中的 bam 文件进行了数据分析，使用 R 软件分别进行了基因表达 counts 矩阵的生成，以及后续的主成分分析、聚类分析、差异基因表达分析。具体分析的程序见网站附件。

结果与讨论

1. 基因表达 counts 矩阵的生成

使用软件： R 语言 Rsamtools 包、GenomicFeatures 包、GenomicAlignments 包，以及 linux 系统的 grep 命令。

具体方法： 首先使用 linux 的 grep 命令从人类基因组注释的全文件中提取了 14 号染色体的基因的注释信息，使用 R 语言的包统计得到了 14 号染色体上共计 1939 个基因的在 8 个 Run 下的 count 数，去掉在八个 Run 中都为 0 的基因，对剩余的 243 个基因进行了分析。

2. 主成分分析

对统计得到的 243 的基因的信息进行主成分分析，并按照前两个累计方差解释率达到 80%的主成分绘制了二维平面图（图 1），如同所示，经过主成分分类，可以看到对照组、两组实验组分离的较好，但两组实验组的区分度不是很好。

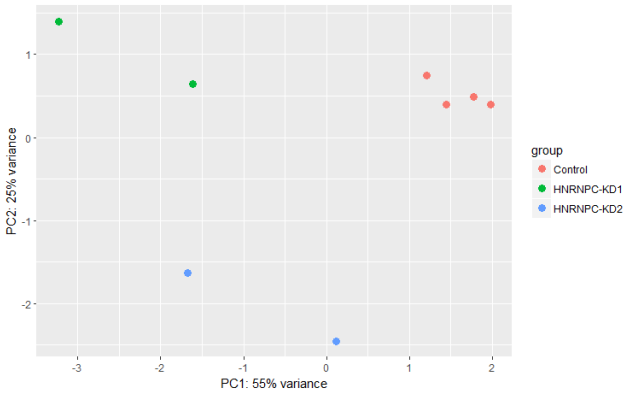


图 1. 不同实验条件下的主成分分析图

3. 差异基因表达分析

为了比较在不同条件下差异基因的表达情况，使用 R 语言的 DEseq2 包进行了分析，绘制了改变倍数最大的基因在三种条件下的表达情况（图 2）。

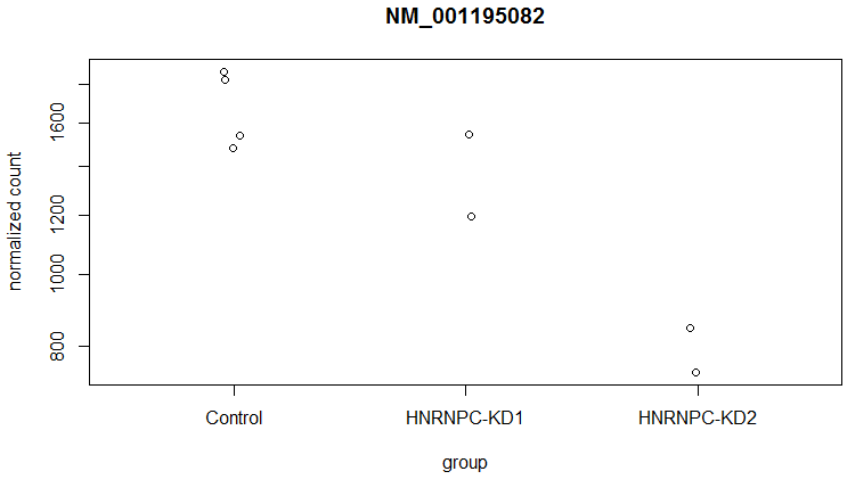


图 2. Foldchange 最大的基因在三种条件下的 count 数

同时绘制了在"HNRNPC-KD1"~"Control"（图 3）、"HNRNPC-KD2"~"Control"（图 4）的 MAplot。并用红色的点区分了变化显著的基因 NM_001195082(*TEX22*)睾丸特异表达。

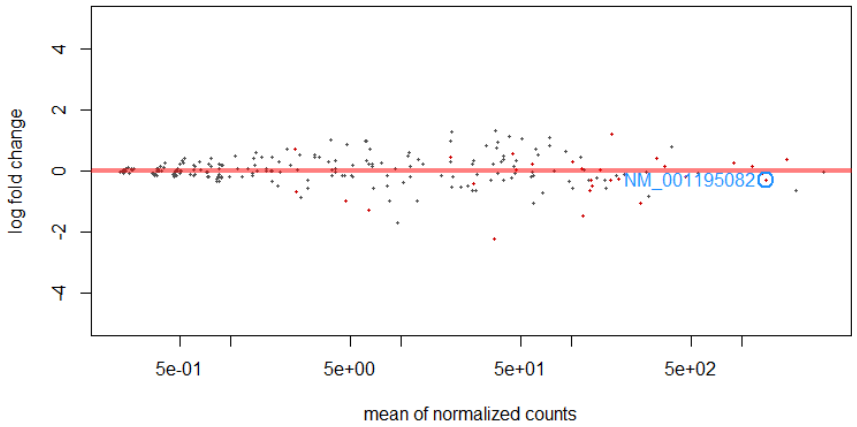


图 3. HNRNPC-KD1.vs.Control

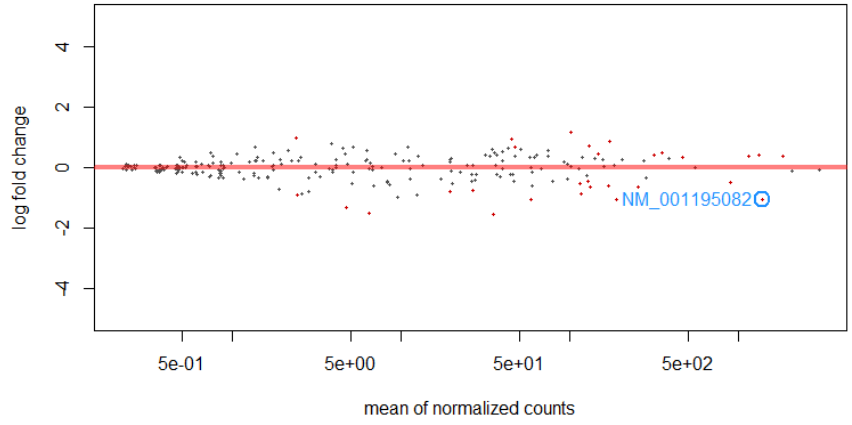
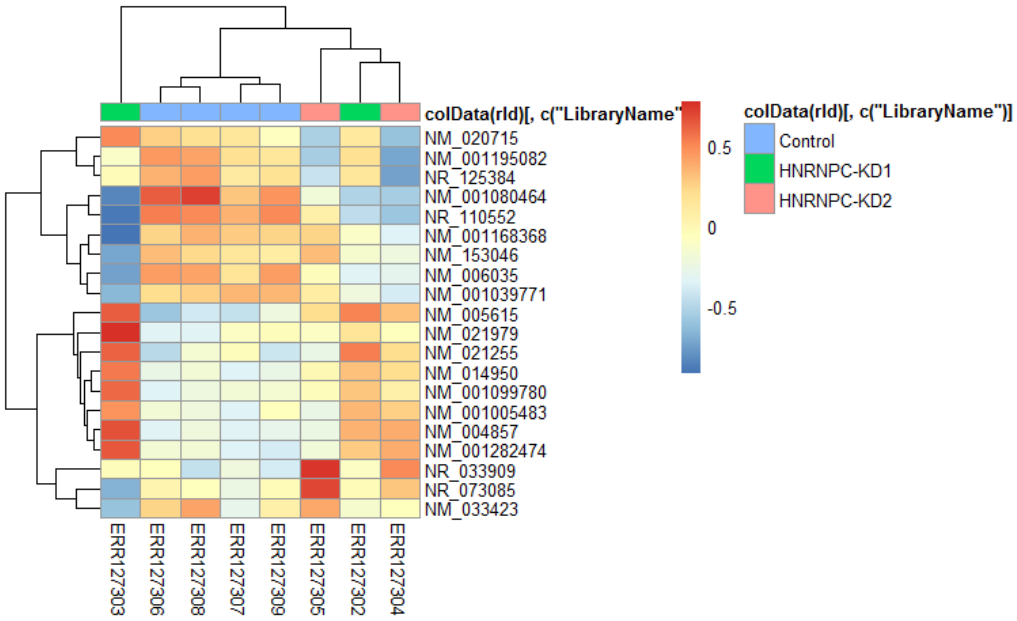


图 4. HNRNPC-KD2.vs.Control

4.聚类分析

使用语言的 DEseq2 包对方差变异在前 20 的基因进行了聚类分析（图 5），从图中的结果发现，使用第一种 siRNA 进行敲除的两组重复间的差异较大，但与对照组相比在一些基因的调控上维持了相同的模式。对照组的重复之间基因表达的模式比较一致。



5. 讨论

综合以上分析结果，第一种 siRNA（HNRNPC-KD1）敲除的实验组的重复之间的差距较大，两组 siRNA 影响的变化最大的基因是 *TEX22* 基因。

参考文献

[1] Zarnack K, König J, Tajnik M, et al. Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements.[J]. *Cell*, 2013, 152(3):453-466.

[2]H. Pages (2017). *RNAseqData.HNRNPC.bam.chr14*: Aligned reads from RNAseq experiment: Transcription profiling by high throughput sequencing of HNRNPC knockdown and control HeLa cells. R package version 0.14.0.