《基因组学数据分析》第二次作业

2023-10-20

考察内容：

1. 掌握基本的 Linux 命令，学习软件的下载与安装。
2. 熟悉 RNA-seq 分析流程，掌握基本操作。
3. 熟悉 RNA-seq 常见图形的绘制。

设备要求：

1. 上游分析需要 Linux 环境，下游分析可在 Rstudio 中完成
2. 参考腾讯云服务器（CPU 4 核，内存 8GB，系统盘 180GB）

作业内容：

一、 简述 RNA-seq 分析流程，以 fastq 文件夹中的数据（8 个*.fastq 文件）为例，完成下列操作：

1. 查看测序文件，统计文件的行数、read 数、测序长度信息。

提示：可能会用到 less、head、wc 命令。

2. 安装软件 fastqc，对测序数据进行质量评估，并对评估结果进行简短的解读，同时附上结果文件中 Basic Statistics 部分的截图（以 SRR1039508 为例）。

提示：推荐用 conda 安装。https://bioconda.github.io/可以查询到一些能用 conda 安装的软件；命令行 conda search 作用类似 。

3. 安装软件 STAR，对压缩包中的参考基因组 Homo_sapiens.GRCh38.dna.chromosome.22.fa 构建索引，并将测序数据（*.fastq）比对到参考基因组上。

要求： 提供构建索引和比对的命令，对参数进行适当解释，给出在 Linux 终端操作的代码或代码截图。以 SRR1039508 为例，提供序列比对率的截图。

提示：22 号染色体相对比较小，STAR 构建索引时需要做参数上的调整，参考 STAR 手册。

4. 根据比对后的结果，对基因表达进行定量。

要求：尝试实现 count 与 TPM 的转换，gene id 与 gene symbol（如 SOX2）的转换，思考如果多个 id 对应同一个 gene symbol 怎么办？将行名为 gene symbol，列包含各样本的 count、TPM 的表达矩阵保存为 csv 文件提交。

提示：可能会用到 gtf 文件、featureCounts 等工具。

二、 请以压缩包中的文件 raw_count.csv 作为输入，完成以下分析：

1. 分析 trt 组和 untrt 组的差异表达基因，画出 treat 组显著（即 padj<0.01）高表达（即 log2FoldChange 大于 1）、低表达（即 log2FoldChange 小于-1）的各 top20（按 log2FoldChange 排序）基因的表达热图，并提供 R 代码。

提示：可以使用 R 的 DESeq2 进行差异分析。

2. 对差异显著（即 padj<0.01 且 log2FoldChange 的绝对值大于 1）的基因进行 GO biological process 富集分析，输出前 5 条通路的结果（作图或者表格形式均可）。

提示：

① 富集分析可用 R 包 clusterProfiler 的 enrichGO 函数，注意基因名称转换；

② 富集分析也可以使用 Gene Ontology Resource（http://geneontology.org/），Enrichr（https://maayanlab.cloud/Enrichr/），WebGestalt（http://www.webgestalt.org/）或 Metascape（https://metascape.org/gp/index.html）等在线资源。

要求：提供富集分析过程中的关键 R 代码或者网页截图。

注：

1. 请将 Linux 代码保存为.txt 或者.sh 文件，R 代码保存为.R 文件，代码和结果进行截图整理到.pdf 文件中。文件前缀为“hw2_YourName_YourStudentID”。所有的文件打包成“作业 2-姓名-学号-学校”压缩文件，教学网用户请在教学网通过教学网=>课程作业提交。非教学网用户发送到邮箱 genomics2023@163.com，邮件主题为《基因组学数据分析》第二次作业；

2. 本次作业占课程总成绩的 20%；

3. 鼓励遇到问题先在网上进行查阅，仍不能解决后再与他人交流学习，但严禁抄袭行为；

4. 请在 2023 年 11 月 10 日晚 24:00 之前提交作业，作业迟交会根据情况酌情扣分。