|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Prompts | Perf. | Alternative Prompts | Perf. | lessons |
|  | Act as an experienced bioinformatician.  Write R code, without explanation, to analyze RNA-Seq read counts data following instructions. |  | Act as an experienced bioinformatician.  Write R code, without explanation, to analyze RNA-Seq read counts data following instructions. Load R packages first. |  | Severely impact performance when specifically asked to load packages. |
| 1  Read data | Read raw counts data from a file called ’../GSE37704\_sailfish.csv’.  Treat the first column as row id. Each row represents a gene.  Print the number of genes and the first 10 rows. | 20% | Read the RNA-seq read counts data from a file called ’../GSE37704\_sailfish.csv’.  Store in a data frame called raw\_counts.  Treat the first column as row id. Each row represents a gene.  The samples are derived from primary human lung fibroblasts.  The first three columns are normal cells while the last three columns are from cells without the Hoxa1 gene.  Print the first 10 rows. | 20% | Unnecessary details degrade performance. |
| 2. Gene annotation | The row names are Ensembl gene IDs. Using the org.Hs.eg.db package, retrieve corresponding gene symbols, Entrez gene IDs, and gene names. Save the information in a CSV file. Print the first 5 rows. | 80% |  |  |  |
| 3  Data transform,  heatmap | Perform the following analyses sequentially.  Calculate counts per million (CPM) from raw counts.  Filter the data by removing genes with CPM less than 1 in all samples.  Print the number of genes removed.  After filtering, add 1 to all CPM numbers.  Then log transform the resultant data using base 2.  Sort genes by standard deviations in descending order.  Convert data as a matrix.  Convert row names to gene symbols using the gene information obtained above. Remove genes without gene symbols.  Subtract row means from all rows.  Create a heatmap of the top 50 genes using blue and yellow colors. | 60% | Calculate counts per million (CPM).  Add 1 to all CPM numbers.  Then log transform using base 2.  Rank genes by standard deviations in descending order.  Subtract row means from all rows.  Create a heatmap of the top 50 genes using blue and yellow colors. |  |  |
| 4 PCA | Continue analysis using the filtered and log-transformed data.  Transpose the data frame and conduct PCA. Use ggplot2 to plot the  samples with the first two components. | 60% | Continue analysis using the log-transformed CPM data.  Transpose the data frame and conduct PCA. Use ggplot2 to plot the  samples using the first two principal components. Add the percentage of variances explained to the axis labels, but only keep one digit after the decimal point. | 20% | Customizing plots makes it harder. |
| 5 DEG | From the raw count data, remove genes not included in the filtered CPM data.  Use the resultant data to identify differentially expressed genes with the DESeq2 package.  The first three columns are control samples.  The last three are hoxa1 knockouts. Remove genes with any missing values in the DESeq2 results.  Generate a volcano plot, with lines at log fold change at 1 and FDR less than 0.01. | 70% | Continue analysis using the raw read counts.  Use DESeq2 to identify differentially expressed genes.  The first three columns are control samples.  The last three are hoxa1 knockout. Convert the result to a data frame.  Generate a volcano plot, with lines at log fold change at 1 and FDR less than 0.01. | 20% | Not removing genes with no change causes errors. |
| 6  Identify  Upregulated,  ID conversion | From the DESeq2 result, identify up-regulated genes with FDR < 0.01 and LFC bigger than 1.  Sort the genes by LFC in descending order.  Using the gene information obtained above, create a data frame containing the following information for the up-regulated genes: log fold change, FDR, gene symbol, Ensembl IDs, Entrez IDs, and gene name.  Save that into a CSV file.  Print the top 10 genes. | 100% |  |  |  |
| 7 | Conduct Gene Ontology enrichment analysis on the upregulated genes using the clusterProfiler package. Focus on GO Biological Process terms,  excluding those with less than 5 or more than 5000 genes. Use the Entrez gene IDs in the last step and annotation in the org.Hs.eg.db package.  Save the results to a CSV file.  Use the enrichplot package to plot the top 10 terms. | 30% |  |  |  |
| 8 | Repeat these analyses for down-regulated genes. | 100% |  |  |  |
| 9 | Use the gage package to retrieve KEGG pathways for humans.  Show the number of pathways. Show the name of the genes in the first pathway. | 30% |  |  |  |
| 10 | Perform pathway analysis using the gage package and the KEGG pathways just retrieved.  Use the log transformed data, but change the row names to Entrez gene IDs using the gene information retrieved. Remove genes without Entrez IDs. The first three samples are controls. The last three are mutant.  Write the pathway analysis results to a file.  Show the information for the top three top upregulated KEGG pathways from the result.  Extract and show the pathway IDs of the three top pathways.  The pathway IDs are the first word of the pathway name. More information about the parameters for the  gage function:  exprs:  an expression matrix or matrix-like data structure, with genes as rows and samples as columns.  gsets a named list, each element contains a gene set that is a character vector of gene  IDs or symbols. For example, type head(kegg.gs). A gene set can also be a  'smc' object defined in PGSEA package. Please make sure that the same gene  ID system is used for both gsets and exprs.  ref:  a numeric vector of column numbers for the reference condition or phenotype  (i.e. the control group) in the exprs data matrix. Default ref = NULL, all columns  are considered as target experiments.  samp:  a numeric vector of column numbers for the target condition or phenotype (i.e.  the experiment group) in the exprs data matrix. Default samp = NULL, all  columns other than ref are considered as target experiments.  set.size:  gene set size (number of genes) range to be considered for enrichment test. Tests  for too small or too big gene sets are not robust statistically or informative biologically. Default to be set.size = c(10, 500).  same.dir:  boolean, whether to test for changes in a gene set toward a single direction (all  genes up or down regulated) or changes towards both directions simultaneously.  For experimentally derived gene sets, GO term groups, etc, coregulation is commonly the case, hence same.dir = TRUE (default); In KEGG, BioCarta pathways, genes frequently are not coregulated, hence it could be informative to let  same.dir = FALSE. Although same.dir = TRUE could also be interesting for  pathways. | 10% |  |  |  |
| 11 | Extract LFC numbers from DESeq2 results. Convert the LFC numbers into a vector, where names of values are Ensembl gene IDs. Convert gene IDs to Entrez using gene information retrieved.  Use the pathview package to generate pathway diagrams for the top three pathways just identified using gage. Color-code the expression data on the diagram using logfc. | 0 |  |  |  |