Statistical Bioinformatics Lab: Even More Advanced R

January 25, 2023

Solution - Mini Activity from Lab 2

Univariate vs. Multivariate

- **Univariate = only 1 variable at a time (i.e. t-test, fold-change)**
  - Describe data, find patterns.
  - Ignores correlations between variables
- **Multivariate = multiple variables analyzed together, considering potential interactions between them (i.e. dimension reduction, regression + clustering)**
  - Multivariate analysis ALWAYS refers to the dependent variable.

Replacing in R (sub, gsub)

- sub() and gsub() are replacement functions, replaces the occurrence of a substring with another substring.
  - sub() - replaces first instance of a substring
  - gsub() - replaces all instances of a substring
- gsub(pattern, replacement, x)

```
> hello <- "Hello my name is Roger. I am a computer."
> sub("a", "A", hello)
[1] "Hello my nAme is Roger. I am a computer."
> gsub("a", "A", hello)
[1] "Hello my nAme is Roger. I Am A computer."
```

Matching in R (match)

- **match()** returns a vector of the position of the first occurence of vector1 in vector2. If no matches, NA returned.
- **match(vector1, vector2)**
- **%in%**

```
> x <- c(1, 2, 3, 4, 5, 6, 7, 8)
> match(7, x)
[1] 7
> 7 %in% x
[1] TRUE
```

Matching in R (which)

- **which()** returns a vector of the position of the all occurrences of a value if it satisfies the specified condition.
- **which(X)**
  - **X is the logical vector**
- Can return multiple matches

```
> x
[1] 1 2 3 4 5 6 7 8
> which(x > 3)
[1] 4 5 6 7 8
> which(x != 1)
[1] 2 3 4 5 6 7 8
```

```
> phyla <- c("Kingdom", "Phylum", "Class", "Order", "Family", "Genus", "Species")
> which(phyla %in% "Genus")
[1] 6
```

Demo

Mini Activity

- Read in the “gene_data.csv” and “sig_genes.csv” files.
- Replace the “_” in the colnames of gene_data with an empty space.
- Match the entrez ids (rownames of gene_data) with only genes from the sig_genes dataset that are significant (adj-pval < 0.1 and logFC > 2).
- Extract gene_data information for only the matches from sig_genes.
- Perform t-tests on all genes between the S and C groups.

75 min