"Tidy" High-throughout Analysis Data, Examplified by RNA Sequencing Data

Tu Hu

19 April 2019

Introduction

High-throughout analysis revolutionized biology and biochemical sciences. Because These disciplines were also brought into a big data generation. Such big data was characterized as high-dimensional (consists hundreds and thousands variables), computing and storage resources demanding (difficult for personal computer to handle the analysis tasks).

Therefore, at the first place, when computing and storage resources were expensive and limited. The data storage mechanisms were designed for fitting the performance of "machine", rather than the habits of "human".

Tidy Format

In a data analysis task, 80% of time is spent on cleaning and preparing the data.

Tidying RNA Sequencing Data

Old fashioned three-table storage

This essay examplified data "tidying" by a subset of GSE5859 gene expression dataset. GSE5859 is a open-source dataset storing gene expression assay results of a US population study(Spielman 2007).

This dataset was transformed to R format by Rafael A. Irizarry for tutorial purpose and available on GitHub.

```
library(devtools)
install_github("genomicsclass/GSE5859Subset")
library(GSE5859Subset)
data(GSE5859Subset)
```

This dataset consists three tables: "geneAnnotation", "geneExpression", "sampleInfo".

```
dim(geneAnnotation)
```

```
## [1] 8793    4
```

```
dim(geneExpression)
```

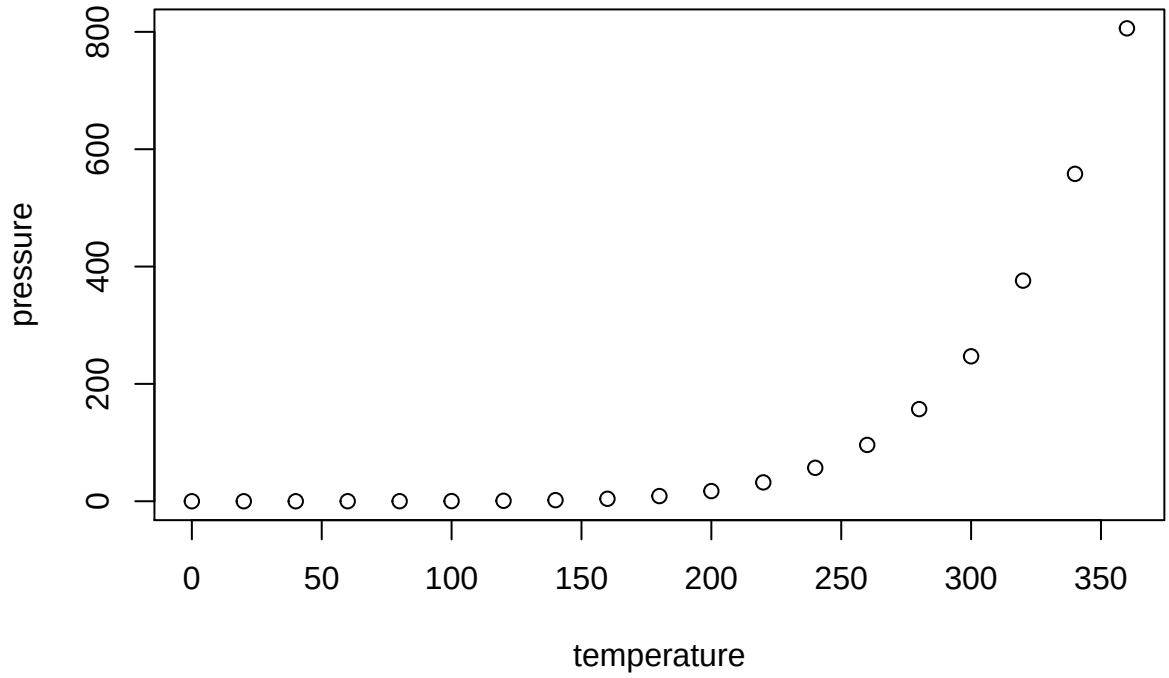
```
## [1] 8793   24
```

```
dim(sampleInfo)
```

```
## [1] 24  4
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

References

Spielman, Richard. 2007. "Common Genetic Variants Account for Differences in Gene Expression Among Ethnic Groups." *Nature Genetics* 39 (7). Nature Publishing Group: 226–31. doi:10.1038/ng1955.