Assignment

Introduction

For this assignment, you will use The Cancer Genome Atlas (TCGA) glioma RNA-sequencing data. This dataset is a matrix of RSEM normalized and log2-transformed gene expression values. This dataset consists of 1,129 samples and was retrived from Broad GDAC Firehose (https://gdac.broadinstitute.org/).

Here, you're tasked to calculate the coefficient of variance (CV2) for each gene. The CV2 is calculated by dividing variance by the square of mean. CV2 is often used as a measurement of variability, and typically genes with moderate-to-high CV2 are brought forward for downstream analysis such as dimension reduction and clustering.

Specifically, you need to benchmark the different approaches taught in *Efficient R Codes* for calculating CV2, namely:

- `for` loops that grow vectors
- `for` loops that don't grow vectors
- Utilizing apply family
- Parallel computing
- Vectorize code

Getting started

```
# Load packages
library(data.table)
library(parallel)
library(microbenchmark)

# Read file
df <- fread("Datasets/GBMLGG.uncv2.mRNAseq_RSEM_normalized_log2.txt", sep="\t",
            header=TRUE, stringsAsFactors=FALSE)

# Check dimensions
dim(df)
```

```
## [1] 18328   702
```

```
# Sneak peek
df[100:105,1:5]
```

```
##             gene TCGA-02-0047-01 TCGA-02-0055-01 TCGA-02-2483-01
## 1: ABHD15|116236        8.118009        8.337514        7.333303
## 2:   ABHD1|84696        3.467749        3.252552        3.510759
## 3:   ABHD2|11057       13.461558       11.877184       11.632199
```

```
## 4:  ABHD3|171586        8.174347        8.393905        7.856536
## 5:   ABHD4|63874       11.328497       10.097301        9.897343
## 6:   ABHD5|51099       10.070969       10.040068        9.099060
##    TCGA-02-2485-01
## 1:        7.330490
## 2:        1.929602
## 3:       11.617205
## 4:        9.109520
## 5:       11.684289
## 6:        9.566235
```