

# Exercises for RNA-seq: Alignment & Counting

March 23, 2015

## Exercise 1

There's an R function called `rowSums()` that calculates the sum of each row in a numeric matrix, like the count matrix we have here, and it returns a vector. There's also a function called `which.max()` that determines the index of the maximum value in a vector. Here's an example of it in action:

### Example

```
# Create a temporary data frame called fake
fake <- data.frame(row.names = c("GeneA", "GeneB", "GeneC"),
                   samp1=c(20,50,40), samp2=c(30,70,50))

# This is what it looks like
fake

##          samp1 samp2
## GeneA      20    30
## GeneB      50    70
## GeneC      40    50

# Get the rowSums
rowSums(fake)

## GeneA GeneB GeneC
##    50   120    90

# Get the index of the maximum total value
which.max(rowSums(fake))
```

```
## GeneB
##      2

# Store that index
topGene <- which.max(rowSums(fake))

# Get that row, and all the columns
fake[topGene, ]

##      samp1 samp2
## GeneB    50    70
```

## Challenge

1. Find the gene with the highest expression across all samples – remember, each row is a gene.
  2. Extract the expression data for this gene for all samples.
  3. In which sample does it have the highest expression?
  4. What is the function of the gene? Can you suggest why this is the top expressed gene?
- 

## Exercise 2

Using the `subset()` function, print out all the columns where the control mean does not equal 0 **and** where the UVB mean does not equal zero.

Bonus: code golf – use the fewest characters to get the same solution.

---

## Exercise 3

1. Plot the mean expression of each gene in control against the UVB sample mean. Are there any outliers?
2. How could you make this plot more informative and look more professional? Hint: try plotting on the log scale and using a different point character.