# EDA of C. elegans dataset

*Domenick Braccia*

*2/26/2019*

## Overview

This document will serve as an exploratory data analysis of data from this publication by Asija Diag *et al.*

### Loading & munching data

Overview of the data present in each .xlsx file:

- S1 : Hermaphrodite expression levels (gonads #1 & #2)
- S2 : Hermaphrodite expression levels (normalized and averaged accross replicates - see M&M)
- S3 : Hermaphrodite dynamic transcript groups
- S4 : Hermaphrodite dynamic transcripts (Profile I)
- S5 : Analysis of comparison with NEXTDB
- S6 : expression levels of X linked genes in hermaphrodites
- S8 : expression levels / transcriptional rates in hermaphrodites
- S9 : Male transcript counts (gonads 1 and 2)
- S10 : Male transcript expression levels
- S11 : transcript to gene name table

```r
# read in necessary libraries
library(readxl)
library(tidyverse)
library(skimr)

### Importing and cleaning tables S1,S2,S8,S9,S10,S11 ###

# importing S1 - hermaphrodite expression table
hermExpNad <- read_excel("../data/S1_Table.xlsx", skip = 3)
colnames(hermExpNad) <- c("ORF", 1:10, 1:10, "WBGene")
glimpse(hermExpNad)
```

```
## Observations: 20,779
## Variables: 22
## $ ORF    <chr> "2L52.1", "2RSSE.1", "2RSSE.2", "3R5.1", "4R79.1", "4R7...
## $ `1`    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ `2`    <dbl> 0, 0, 0, 7, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ `3`    <dbl> 0, 0, 0, 12, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, ...
## $ `4`    <dbl> 0, 0, 0, 16, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ `5`    <dbl> 0, 0, 0, 9, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ `6`    <dbl> 2, 0, 0, 9, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2...
## $ `7`    <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ `8`    <dbl> 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3...
## $ `9`    <dbl> 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3...
## $ `10`   <dbl> 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1...
## $ `1`    <dbl> 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1...
## $ `2`    <dbl> 0, 0, 0, 9, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ `3`    <dbl> 0, 0, 0, 9, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ `4`    <dbl> 0, 0, 0, 8, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1...
## $ `5`    <dbl> 0, 0, 0, 9, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 7...
## $ `6`    <dbl> 1, 0, 0, 6, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 8...
## $ `7`    <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 8...
## $ `8`    <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3...
## $ `9`    <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 7...
## $ `10`   <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 2...
## $ WBGene <chr> "WBGene00007063", "WBGene00007064", "WBGene00044165", "...
```

```
# S2 Herm exp levels;normalized and averaged
hermExpNormed <- read_excel("../data/S2_Table.xlsx", skip = 3)
colnames(hermExpNormed)[1] <- "ORF"
colnames(hermExpNormed)[17] <- "Gene_Name"
glimpse(hermExpNormed)
```

```
## Observations: 20,687
## Variables: 18
## $ ORF        <chr> "2L52.1", "2RSSE.1", "2RSSE.2", "3R5.1", "4R79.1", ...
## $ `1`        <dbl> 0.0000000, 0.0000000, 0.0000000, 1.1285917, 0.00000...
## $ `2`        <dbl> 0.000000, 0.000000, 0.000000, 4.934989, 0.000000, 0...
## $ `3`        <dbl> 0.0000000, 0.0000000, 0.0000000, 3.9633200, 0.00000...
## $ `4`        <dbl> 0.0000000, 0.0000000, 0.0000000, 3.4258657, 0.00000...
## $ `5`        <dbl> 0.0000000, 0.0000000, 0.0000000, 2.2290508, 0.00000...
## $ `6`        <dbl> 0.5280138, 0.0000000, 0.0000000, 2.7463885, 0.00000...
## $ `7`        <dbl> 0.9536161, 0.0000000, 0.0000000, 0.4751858, 0.00000...
## $ `8`        <dbl> 0.000000, 0.000000, 0.000000, 2.373090, 0.000000, 0...
## $ `9`        <dbl> 0.0000000, 0.0000000, 0.0000000, 1.4237781, 0.00000...
## $ `10`       <dbl> 0.000000, 0.000000, 0.000000, 1.197841, 0.000000, 0...
## $ Chromosome <chr> "CHROMOSOME_II", "CHROMOSOME_II", "CHROMOSOME_II", ...
## $ Start      <dbl> 1867, 15268136, 15274315, 13780127, 17486946, 17480...
## $ End        <dbl> 4663, 15273238, 15275613, 13781032, 17490461, 17483...
## $ Strand     <chr> "+", "+", "+", "+", "-", "-", "+", "+", "+", "-", "...
## $ WBGene     <chr> "WBGene00007063", "WBGene00007064", "WBGene00044165...
## $ Gene_Name  <chr> NA, NA, NA, "pot-3", "nas-6", NA, NA, "sri-20", "sp...
## $ Type       <chr> "cds", "cds", "cds", "cds", "cds", "cds", "cds", "c...
```

```
# S8
hermRates <- read_excel("../data/S8_Table.xlsx", skip = 3)
colnames(hermRates) <- c("ORF", "Gene_Name", "WormBase_ID", "Chromosome", "Strand", "Ribo_rep1", "Ribo_
glimpse(hermRates)
```

```
## Observations: 19,977
## Variables: 9
## $ ORF         <chr> "B0019.1", "B0019.2", "B0025.1a", "B0025.2", "B002...
## $ Gene_Name   <chr> "amx-2", NA, "vps-34", "csn-2", NA, NA, NA, NA, "a...
## $ WormBase_ID <chr> "WBGene00000138", "WBGene00007094", "WBGene0000693...
## $ Chromosome  <chr> "I", "I", "I", "I", "I", "I", "I", "I", "I", "I", ...
## $ Strand      <chr> "-", "+", "+", "+", "+", "-", "+", "-", "+", "+", ...
## $ Ribo_rep1   <dbl> 6.5816939, 5.1639885, 13.8514375, 70.1363533, 19.9...
## $ Ribo_rep2   <dbl> 4.4000493, 4.8461051, 13.9819546, 100.0447569, 34....
## $ mRNA_rep1   <dbl> 30.1078502, 86.9022041, 119.9181990, 321.2644476, ...
## $ mRNA_rep2   <dbl> 18.9364571, 120.9614503, 137.9656158, 365.9760174,...
```

```
# S9
maleExpNad <- read_excel("../data/S9_Table.xlsx", skip = 3)
colnames(maleExpNad) <- c("ORF", 1:10, 1:10, "WBGene")
```

```
glimpse(maleExpNad)
```

```
## Observations: 20,779
## Variables: 22
## $ ORF    <chr> "2L52.1", "2RSSE.1", "2RSSE.2", "3R5.1", "4R79.1", "4R7...
## $ `1`    <dbl> 0, 0, 0, 6, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ `2`    <dbl> 0, 0, 0, 8, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ `3`    <dbl> 0, 0, 0, 14, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ `4`    <dbl> 0, 0, 0, 3, 0, 0, 0, 0, 2, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ `5`    <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 19, 5, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ `6`    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 10, 4, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ `7`    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 13, 1, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ `8`    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 6, 0, 2, 0, 0, 0, 0, 0, 0, 1, 2...
## $ `9`    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ `10`   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ `1`    <dbl> 0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ `2`    <dbl> 0, 0, 0, 8, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ `3`    <dbl> 0, 0, 0, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ `4`    <dbl> 0, 0, 0, 7, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ `5`    <dbl> 0, 0, 0, 6, 0, 0, 0, 0, 14, 4, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ `6`    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 13, 5, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ `7`    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 9, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ `8`    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 15, 2, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ `9`    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 6, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ `10`   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ WBGene <chr> "WBGene00007063", "WBGene00007064", "WBGene00044165", "...
```

```
# S10
maleExpNormed <- read_excel("../data/S10_Table.xlsx", skip = 3)
colnames(maleExpNormed)[1] <- "ORF"; colnames(maleExpNormed)[colnames(maleExpNormed) == "Gene Name"] <-
glimpse(maleExpNormed)
```

```
## Observations: 20,687
## Variables: 19
## $ ORF        <chr> "2L52.1", "2RSSE.1", "2RSSE.2", "3R5.1", "4R79.1", ...
## $ `1`        <dbl> 0.000000, 0.000000, 0.000000, 4.265513, 0.000000, 0...
## $ `2`        <dbl> 0.0000000, 0.0000000, 0.0000000, 5.2675156, 0.00000...
## $ `3`        <dbl> 0.0000000, 0.0000000, 0.0000000, 4.8864558, 0.00000...
## $ `4`        <dbl> 0.0000000, 0.0000000, 0.0000000, 3.1359046, 0.00000...
## $ `5`        <dbl> 0.000000, 0.000000, 0.000000, 1.848302, 0.000000, 0...
## $ `6`        <dbl> 0.0000000, 0.0000000, 0.0000000, 0.0000000, 0.00000...
## $ `7`        <dbl> 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0...
## $ `8`        <dbl> 0.0000000, 0.0000000, 0.0000000, 0.0000000, 0.00000...
## $ `9`        <dbl> 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0...
## $ `10`       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ diffExp    <dbl> 0.0000000, 0.0000000, 0.0000000, 2.6478937, 0.00000...
## $ Chromosome <chr> "CHROMOSOME_II", "CHROMOSOME_II", "CHROMOSOME_II", ...
## $ Start      <dbl> 1867, 15268136, 15274315, 13780127, 17486946, 17480...
## $ End        <dbl> 4663, 15273238, 15275613, 13781032, 17490461, 17483...
## $ Strand     <chr> "+", "+", "+", "+", "-", "-", "+", "+", "+", "-", "...
## $ WBGene     <chr> "WBGene00007063", "WBGene00007064", "WBGene00044165...
## $ Gene_Name  <chr> NA, NA, NA, "pot-3", "nas-6", NA, NA, "sri-20", "sp...
## $ Type       <chr> "cds", "cds", "cds", "cds", "cds", "cds", "cds", "c...
```

```
# S11
txrptGene <- read_excel("../data/S11_Table.xlsx")
glimpse(txrptGene)
```

```
## Observations: 5,633
## Variables: 2
## $ geneNames <chr> "3R5.1", "AH6.5", "B0001.1", "B0001.2", "B0001.3", "...
## $ wbgene    <chr> "WBGene00007065", "WBGene00003231", "WBGene00003010"...
```

To this point, I have imported and cleaned up most of the data I think we need. Some further steps may need to be taken to process data, but that can be done on an as needed basis.