Proteomics

BST 226
Statistical Methods for Bioinformatics
David M. Rocke

Chicken Proteomics Data

- "Chicken Corneocyte Cross-Linked Proteome," Robert H. Rice, Brett R. Winters, Blythe P. Durbin-Johnson, and David M. Rocke, *J. Proteome Res.*, 2013, 12 (2), pp 771–776.
- 32 samples
- Four chickens
- Beak, Claw, Feather, Scale
- Soluble fraction, insoluble fraction
- 224 identified proteins

Chicken.Proteomics.xslx

Column	Content
1 (A)	Row Number (Obs)
2 (B)	Protein
3 (C)	Accession Number
4 (D)	MW in kDa
5–36 (E–AJ)	Sample number 1–32

Chicken.Factors.txt

- 32 by 3 table with a header row
- Columns are Chicken (1-4), Component (Beak, Claw, Feather, Scale), and Fraction (Soluble, Insoluble).

Data processing

- Read Chicken.Proteomics.xsls into R
  - You could save the sheet as a tab delimited text file and read with read.delim()
  - You could save the sheet as a .csv file and read with read.csv()
  - You could install and load the CRAN library xlsx
    - You need to install java for this to work
    - chick <- read.xlsx("Chicken.Proteomics.xlsx", sheetIndex = 1, stringsAsFactors = F)
- Extract columns 5–36 as a matrix of counts

Data processing

- Read Chicken.Factors.txt using read.delim().
- Replace the "Chicken" variable with a factor (otherwise "Chicken" is a number).
- Omit proteins (rows of the data matrix) where the number of zeroes is too high (for example, greater than 75% of the sample size, which is (0.75)(32) = 24.
- For each remaining protein, fit a glm model with the quasipoisson family and main effects only using the three variables.

Data processing

- For each remaining protein, fit a glm model with the quasipoisson family and main effects only using the three variables.
- Save the p-values for component and fraction.
- Use p.adjust() to get the FDR adjusted p-values, and select the proteins that have FDR adjusted p-values less than 0.10 for either component or fraction.
- Install the package multcomp, which allows the use of post-hoc tests for many different regression-type estimators.

```
require(xlsx)
require(multcomp)
chick <- read.xlsx("Chicken.Proteomics.xlsx",sheetIndex=1,stringsAsFactors=F)
chickmat <- chick[,5:36]
vars <- read.delim("Chicken.Factors.txt")
vars$Chicken <- as.factor(vars$Chicken)
print(table(apply(chickmat==0,1,sum)))
rowzeroes <- apply(chickmat==0,1,sum)
minpos <- 24
chickmat <- chickmat[rowzeroes <= minpos,]

glms <- function(mat,vars)
{
  m <- nrow(mat)
  pvmat <- matrix(rep(0,3*m),ncol=3)
  for (i in 1:m)
  {
    new.dat <- data.frame(vars,t(mat[i,]))
    names(new.dat)[4] <- "y"
    gobj <- glm(y~Chicken+Component+Fraction,data=new.dat,family=quasipoisson)
    pvs <- drop1(gobj,test="Chisq")[-1,4]
    pvmat[i,] <- pvs
  }
  return(pvmat)
}
```