Cluster Analysis of Wasps

Lindsay Rutter

April 25, 2016

Introduction

```
> #%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
> #%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
> #################################### EXTRA DATA QUALITY####################################
>
> rm(list=ls())
> load("All_wasp.rda")
> listcond = rep(c("DR","DU","F","NR","NU"),each= 6)
> # 157,691 genes
> y = DGEList(counts=countTable, group=listcond)
> keep <- rowSums(cpm(y)>1) >= 6
> # it seems library sizes are recalculated (y$samples$lib.size = colSums(y$counts))
> y <- y[keep, keep.lib.sizes=FALSE]
> ######################## EXTRA FILTERING AT THIS STEP #######################
>
> RowSD = function(x) {
+   sqrt(rowSums((x - rowMeans(x))^2)/(dim(x)[2] - 1))
+ }
> yt = y
> yt2 = as.data.frame(yt[[1]])
> y = mutate(yt2, mean = (DR.1+DR.2+DR.3+DR.4+DR.5+DR.6+DU.1+DU.2+DU.3+DU.4+DU.5+DU.6+F.1+F.2+F.3+F.4+F.
> rownames(y)=rownames(yt)
> # The first quartile threshold of mean counts across the 12 samples
> q1T = as.numeric(summary(y$mean)["1st Qu."])
> # 24,610 genes
> d2q1 = subset(y,mean>q1T)
> # The first quartile threshold of standard deviation across the 12 samples
> q1Ts = as.numeric(summary(d2q1$stdev)["1st Qu."])
> # 18,458 genes
> d2q1 = subset(d2q1,stdev>q1Ts)
> # 14,355
> filt = subset(y,mean<=q1T|stdev<=q1Ts)
> model = loess(mean ~ stdev, data=d2q1)
> # 11,998 genes
> d2q1 = d2q1[which(sign(model$residuals) == 1),]
> d2q1 = d2q1[,1:(ncol(d2q1)-2)]
> # (filt 14,355 genes)
> filt = filt[,1:(ncol(filt)-2)]
> colnames(filt)=colnames(d2q1)
> # filt (30,670 genes)
> filt = rbind(filt,d2q1[which(sign(model$residuals) == -1),])
> #filts = t(apply(as.matrix(filt), 1, scale))
```

```
> #colnames(filts)=colnames(d2q1)
> colnames(filt)=colnames(d2q1)
> y = DGEList(counts=d2q1, group=listcond)
> y = calcNormFactors(y)
```

```
> ggparcoord(data.frame(y[[1]]), columns=1:ncol(y[[1]]), alphaLines=0, boxplot=TRUE, scale="globalminmax
```

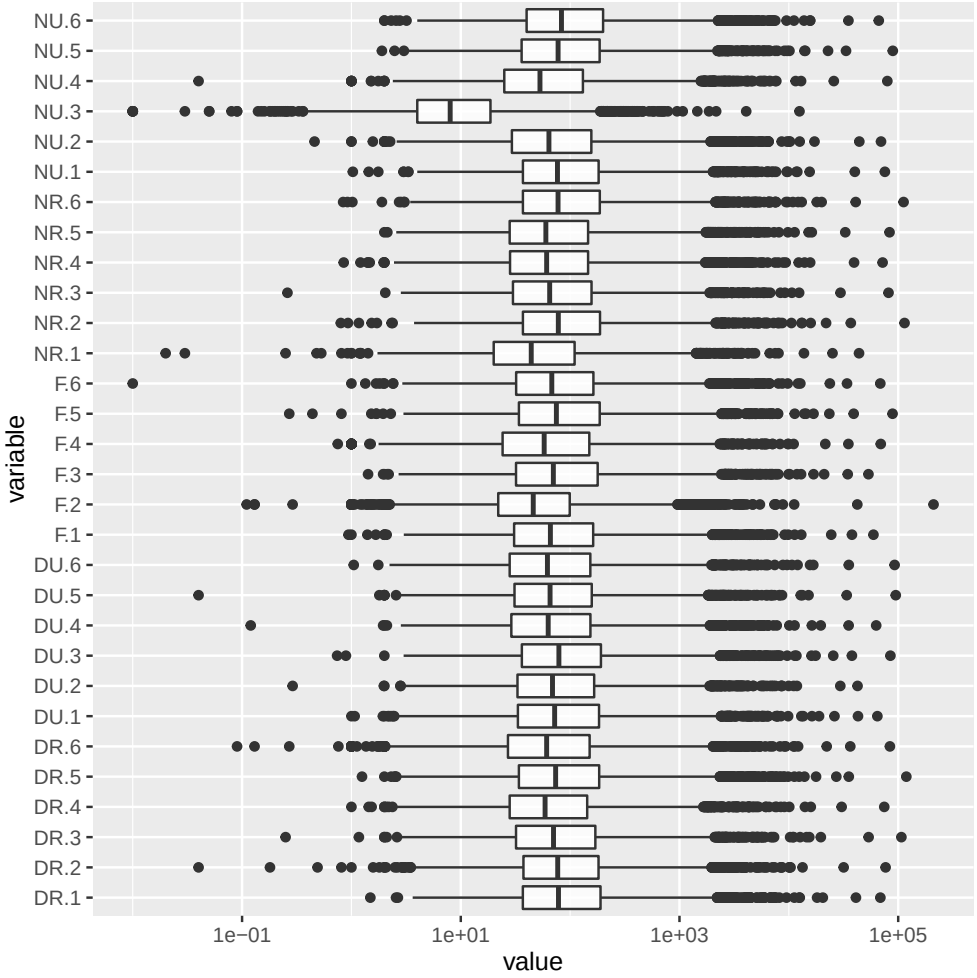

Figure 1: This is all data filtered on CPM and filtered by Loess, then normalized.

```
> ggparcoord(data.frame(y[[1]]), columns=1:6, scale="globalminmax", alphaLines = 0.07)
```

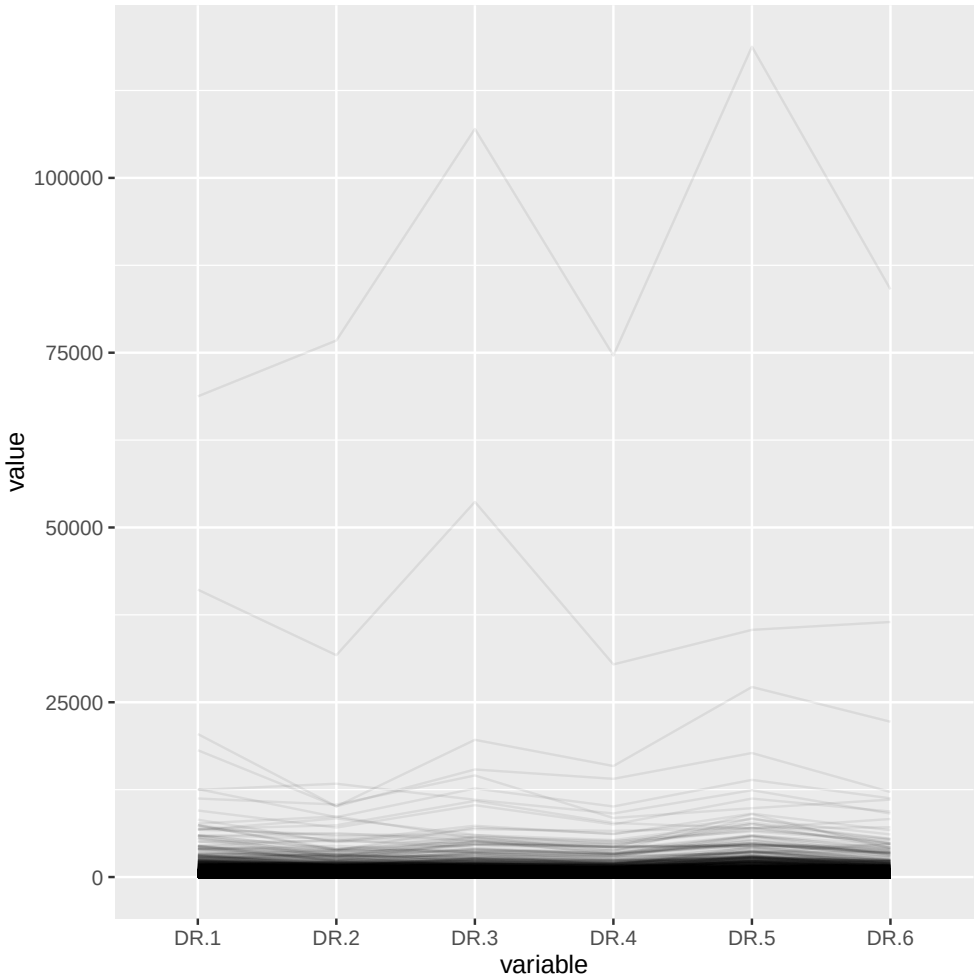

Figure 2: DR PCP

```
> ggparcoord(data.frame(y[[1]]), columns=7:12, scale="globalminmax", alphaLines = 0.07)
```

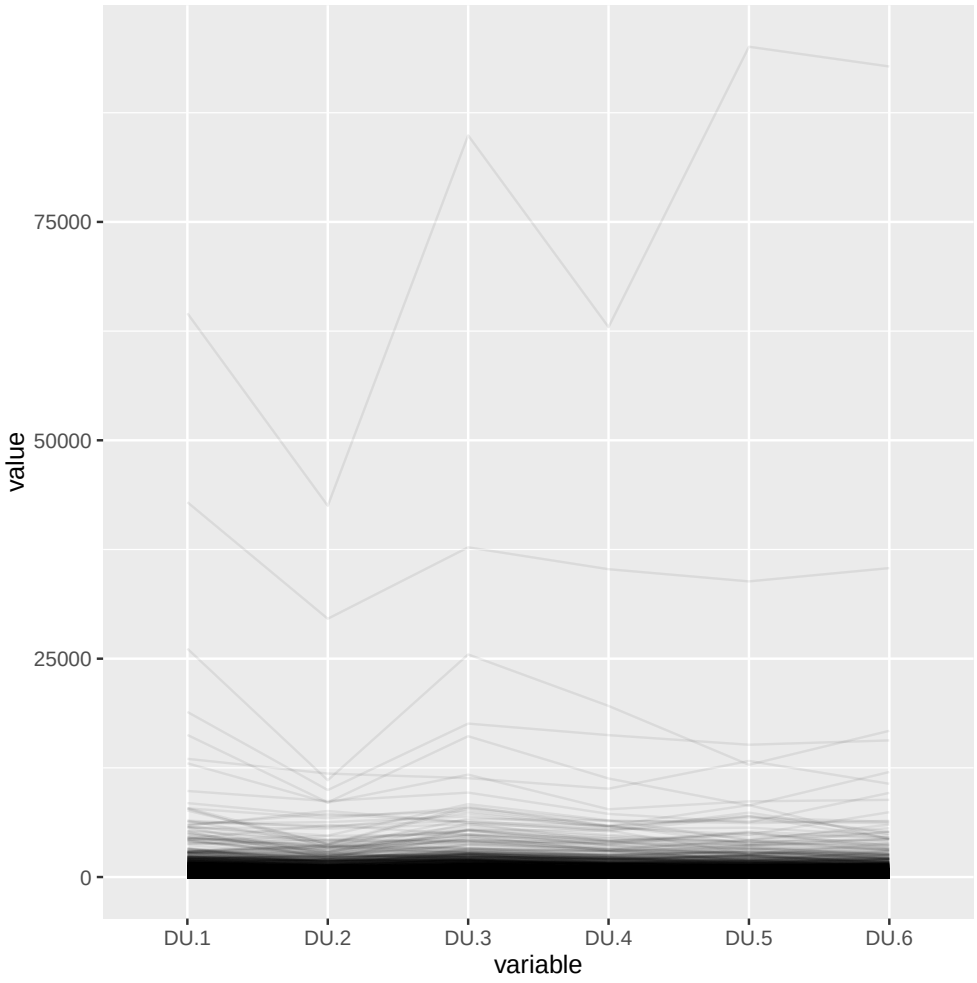

Figure 3: DU PCP

```
> ggparcoord(data.frame(y[[1]]), columns=13:18, scale="globalminmax", alphaLines = 0.07)
```

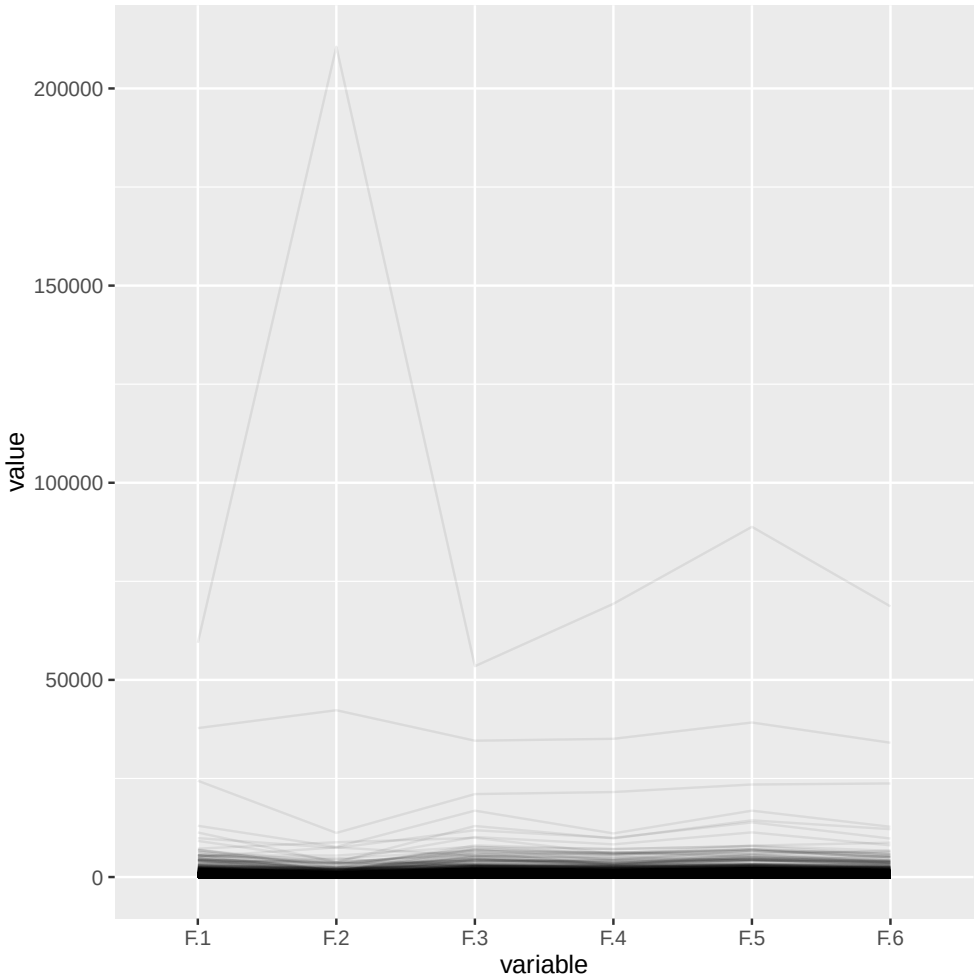

Figure 4: F PCP

```
> ggparcoord(data.frame(y[[1]]), columns=19:24, scale="globalminmax", alphaLines = 0.07)
```

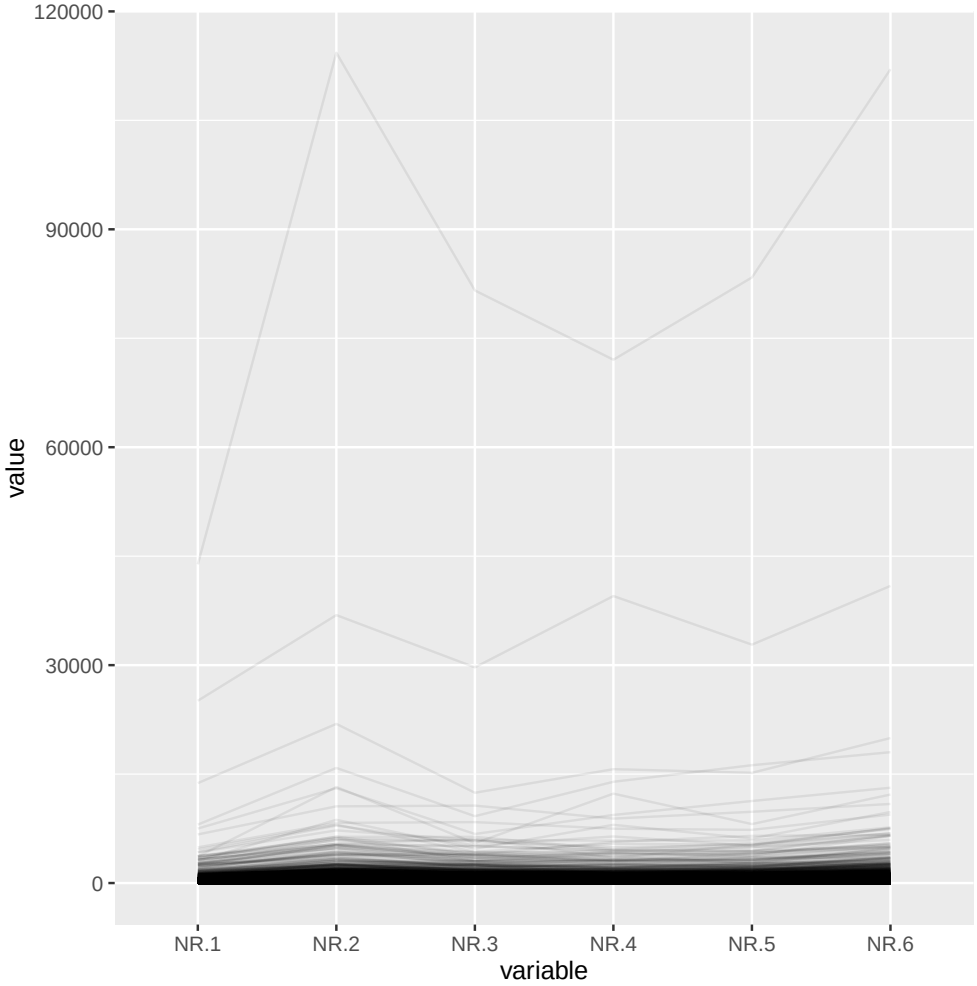

Figure 5: NR PCP

```
> ggparcoord(data.frame(y[[1]]), columns=25:30, scale="globalminmax", alphaLines = 0.07)
```

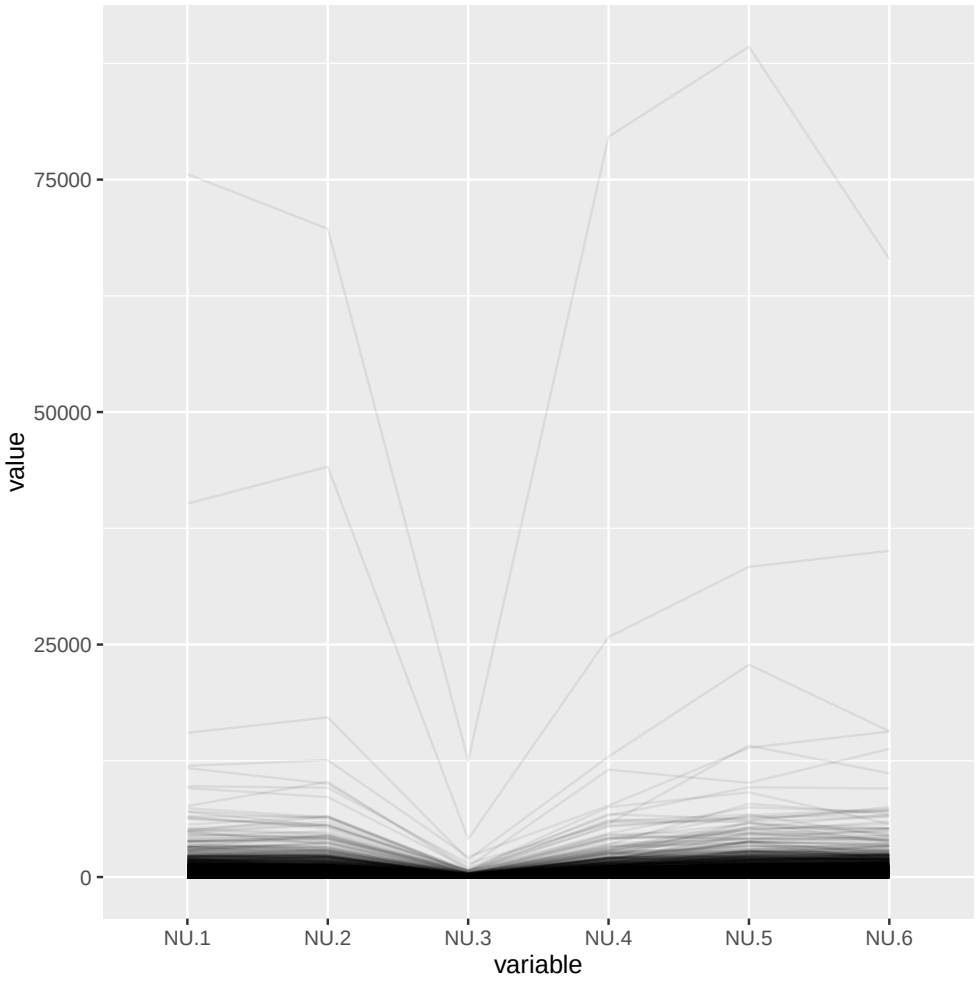

Figure 6: NU PCP